DOCUMENT RESUME

ED 113 684 CS 002 187

AUTHOR Popp, Helen M.; Porter, Douglas
TITLE Measuring the Readability of Children's Trade Books.
INSTITUTION Harvard Univ., Cambridge, Mass. Graduate School of Education.
SPONS AGENCY Ford Foundation, New York, N.Y.
PUB DATE Jul 75
NOTE 133p.

EDRS PRICE MF-$0.76 HC-$6.97 Plus Postage
DESCRIPTORS *Childrens Books; *Childrens Literature; Cloze Procedure; Individual Reading; *Measurement Instruments; Primary Education; *Readability; Reading Comprehension; *Reading Level; Reading Materials; Reading Material Selection; Reading Skills

ABSTRACT

In order to utilize interesting children's trade books in a systematic reading program, two readability formulas were devised based on a selection of children's trade books. Children's scores on selections from these books and judges' rankings were compared. The judges' decisions were considered to be highly credible and were used as the criterion measure. Correlations are reported between 23 textual variables and the criterion. Two readability formulas based on these variables are discussed: one based on a regression analysis that included publisher and linguistic variables (Formula P) and one that included only the linguistic variables (Formula L). Formula P uses words per page and impersonal pronouns plus third person personal pronouns per sentence in the regression equation and correlates .819 with the judges' summed rankings on 50 books. Formula L uses words over seven letters long per sentence and impersonal pronouns plus third person personal pronouns per sentence in the regression equation and correlates .718 with reading difficulty as determined by judges' rankings on 50 books. The books used are listed by title and author, and the judges summed rankings and the regression formulae rankings are given. Grade level scores were not considered appropriate to be assigned. (MKM)

Measuring the Readability of Children's Trade Books

by

Helen M. Popp and Douglas Porter
Harvard Graduate School of Education

July 1975

Acknowledgements

The authors gratefully acknowledge the contributions of many individuals and institutions to this research. The most important contributions were made by the children who read books and told us what they thought. The children, teachers, and administrators of the Somerville, Massachusetts, Public Schools contributed facilities, cooperation, and assistance that we hope will be returned in future years. We also benefited from the assistance and cooperation of the following: Arlington, Massachusetts, Public Schools; The Friends School of Cambridge, Massachusetts, and the Boston Summer School, West Roxbury, Massachusetts.

It was a joy to work with the research assistants from Harvard and Radcliffe Colleges, the Harvard Graduate School of Education, and Goddard College, who did the really hard work.

We also wish to acknowledge with appreciation the reactions to earlier drafts of this report. Professor Jeanne Chall was especially helpful.

Finally, we want to acknowledge the support of The Fund for the Advancement of Education of the Ford Foundation, and especially Marjorie Martus, Program Officer. At the time of our grant (1966-68) the Foundation was active in the field of reading research and we feel that they have been responsible for a great deal of significant research in reading.

TABLE OF CONTENTS

PAGE

APPENDICES (continued)

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

The Problem

A long-range problem in reading instruction has been to provide children who have "cracked the code" with an interesting and systematic set of reading materials that will both teach and encourage further reading. An ideal set of materials would start with the simplest levels of "natural" language, that is, language not written-down or stilted in expression as is the case with much that is written in textbooks, and gradually progress to levels of difficulty and topics that approach what the reader will find in the world around him. There would be a variety of topics to suit the varied tastes of young readers, and there would be some way of matching readers and materials. Children's "trade" books present an interesting and attractice universe of reading matter that might be so used, were it systematized.

There are three major tasks involved in making systematic use of children's trade books: scaling the books for reading difficulty; specifying the necessary prerequisite or entry behavior a child must have in order to read the materials; and, devising instruments and procedures for matching children to the materials. The research described here is concerned with the scaling problem for books for children in the primary grades. However, in the process of developing a readability measure for trade books, procedures and preliminary instruments for assessing children's reading skills vis-a-vis the scaled books had to be devised. Those procedures and instruments are also described in this report.

There are, of course, many extant readability formulas (see Klare, 1963 and 1974). However, they were designed for and validated upon other sorts of reading matter than young children's trade books. For example, the Dale-Chall formula (Dale & Chall, 1948) is not applicable to the very early levels of difficulty, and the Spache formula, (Spache, 1953) which covers the appropriate difficulty levels, was devised for and validated with primary textbooks. It remains to be seen whether a measure devised specifically for trade books will be able to explain more of their variance in difficulty than one of the existing formulas. The readability measures reported here make it feasible to investigate this problem.

Research Strategy

For a number of reasons, we decided to devise a readability measure from the ground up, rather than depending upon criterion data or variables used in past research:

1) Trade books had not been dealt with as a distinct universe of reading matter in past research.

2) The language, format, and illustrations of trade books appeared sufficiently different from textbooks that different variables might be crucial in measuring their difficulty.

3) An appraisal of past work shows that most readability measures have not been based upon a direct performance criterion. That is, most readability formulas have not been developed and validated upon data gathered by having children read samples of the materials to be scaled. Strangely, but understandably, this has been the case especially for those measures designed for the easiest reading levels. It is no mean task to gather a suitable amount of systematic reading performance data

from very young children. Both to rectify the general lack of a performance criterion for readability measures for children's books, and because we believed that a performance criterion would give the most valid basis for developing a new measure, we decided to develop a criterion based upon young children's performances in reading a sample of trade books. As will be seen below, methodological problems suggested the use of a non-performance criterion in addition to the primary, performance criterion.

Overview of Methods and Procedures

The basic procedures were to gether a sample of trade books; have them read by children to produce data for calibrating difficulty of the books; record and analyze selected textual variables that might reveal difficulty of the books; and, finally, using linear regression analysis, devise regression formulas to predict reading difficulty of the books.

The following list of steps is intended to give the reader a detailed grasp of our overall methodology and to guide one to those areas of the report of greatest interest.

1. Sample the universe of children's trade books.

From the available resources we identified and selected a corpus of books, intended to represent the universe of children's trade books (see Part II).

2. Select children to screen books.

Children from a wide range of socioeconomic, ethnic, and public/private school groups (see Part II) read the books prior to final selection. The target group of readers were primarily low SES children (see Part V).

3. Prepare several reasonably equivalent sets of books, selected from the above corpus.

Books were placed into sets of twenty, judged to be equivalent in difficulty, range, interest, attractiveness, and suitability for use in obtaining criterion data (see Part III).

4. Devise procedures for matching reading ability of individual children to appropriate difficulty levels of the books.

Placement tests were designed that permitted children to begin the reading task at a level where they could be successful, and move in a controlled fashion to successively more difficult and easier books (see Part III).

5. Develop criterion measures.

Several alternative criterion measures were selected on rational grounds, and subjected to experimental trial, using data gethered from the above placement tests. A modified cloze comprehension testing criterion was adopted on the basis of data gathered (see Part IV). Oral error scores were also recorded. An alternative criterion measure, based upon expert judgment of book difficulty was also developed (see Part V).

6. Select children to serve as a target group of readers of the books (see Part V).

7. Criterion data gathered.

Using the above sets of books, placement tests, and criterion measures, criterion data were gathered, adjusted to a common baseline level of performance, and combined into a difficulty score for each book (see Part V).

8. Develop and refine readability predictor variables.

Through analysis of past research, examination of the sets of trade books, and by a series of independent correlational analyses, a set of potential predictor variables was isolated.

9. Regression analyses performed.

Several regression analyses were performed to pick alternative readability formulas, dependent upon different base variables which might be preferable under different circumstances (see Part VII).

General Methodological Consideration

The traditional linear regression model was used because of its convenience for computer analysis and the general unavailability of non-linear techniques. Because of the wide range of difficulty and character of the textual materials, from the very simplest prose to relatively difficult, "impressionistic" writing, we expected that some predictor variables would operate over only a part of the range of difficulty and/or change the slope of their correlation with the criterion at some point in the difficulty continuum. Given the possibility of such non-linear variables, and linear regression procedures, one must recognize the possibility that our analyses may have omitted some potentially strong variables.

An early decision was made to use computer analysis of the textual variables, as well as for the regression analyses. This choice was influenced by two major factors: we anticipated a great deal of analysis of many samples of text, far beyond what could be handled reasonably by any but the most ascetic, scholarly monks; then, we wished to make relatively rapid and routine readability analyses available to others, and believed that one way of ensuring this was through the use of easily shared computer programs. A corollary decision was to

consider only those variables that could be key-punched for computer input without coding, that is, only those variables the computer could recognize in the form of natural text input. The result of these decisions has been to eliminate some variables interesting from a theoretical viewpoint, and to give the research an applied rather than basic character. The result is felt most strongly in the elimination of some variables that tap the structural complexity of writing (e.g., dependent clauses, etc.), as opposed to those that indicate level of vocabulary and related facets of the intellectual level of discourse. Basically, the computer procedures used could count words in predetermined categories but could not tag and count structures, such as prepositional phrases, that occur as instances of an open-ended set.

As can be appreciated, the most demanding, difficult, and interesting aspect of this work has been the gathering of suitable, systematic criterion data. In the process of calibrating the difficulty of 60 books, 197 children have read 1072 selections, producing over 400 hours of taped material that was scored for reading errors. Because we wished to expand the data base without gathering further error data, another twenty books were added and judgmental data were used to scale these twenty and the original sixty books as one set of eighty books. These data proved to be extremely stable and systematic, and correlated very highly with the primary criterion of reading errors. Because of within book sampling variability and other unsystematic features of the primary criterion data, these judgmental data have played an important role in developing the readability measures reported below. This aspect of the research is described further in Parts V and VII.

II. SELECTION OF BOOKS

We wanted whatever system was eventually devised for scaling the books to be generalizeable to other popular trade books available to children. For that reason, it was necessary to collect some systematic subset of the enormous number of available books. The primary consideration was that the selected books had to be those which children would indeed want to read; but individual children's interests are varied and may be influenced by many factors. While it is helpful to know that primary children, in general, may prefer fairy tales and stories about animals, one cannot assume that guidance from such generalizations about children's interests will suit all individuals (Neumeyer, 1968). Therefore, our goal was to collect a set of materials which would reflect the accumulation of published children's literature and thus include a wide range of different subject matter and variations in vocabulary, syntactic structure, and style. Using such a varied set for our study, we assumed, would allow generalizations to an extensive set of books from which children would be able to make selections for their own pleasure. There follows a description of the procedures used in acquiring books, and then, from those acquired, selecting the ones to be included in the study sample.

Books were purchased at local bookstores and school and library supply houses by a heterogeneous group of undergraduates from several colleges, students at Harvard Graduate School of Education, and project research assistants. Over a period of 18 months these eight individuals selected books which they thought would be appealing to children in the primary grades. They looked for books ranging from the very simplest

to those they thought could be read by eight or nine-year olds who were excellent readers. They referred to current book reviews, spoke with children's librarians, and sought specific suggestions from their very young friends. There were no restrictions as to size, shape, or cost of the books. It was agreed that selections would have to be read by children in order to gain evidence of their appeal and that each would have to defend his selections and answer whatever criticisms might occur.

Approximately 170 books were purchased concurrently with children's trial readings and discussions of the relative merits of the books. Not all books were equally acceptable to the entire group of selectors and deletions from the collection occurred as the researchers observed second and third grade children reading them at various sites (an upper SES private school in Cambridge, a middle SES public school in Arlington, and a lower SES summer-school in West Roxbury). During this time, two procedures were used to gather information on the appeal of the selected books to children. First, about 50 children from different neighborhoods were interviewed concerning the books, and then, the children read the books out loud.

Four researchers went out in pairs for the initial interviews with first, second, and third graders in two nearby schools. Children were taken from their classrooms to a room assigned for our use. Each child was told that we were gathering a collection of books for a library and that we needed their help in determining which ones he would like to read or have read. Each was told that he could read parts of any book to himself. We also asked the children to read short passages aloud to us and we, in turn, read to them. One researcher of the team questioned the children in a non-directive way while the other

researcher taped the children's comments and noted which books were approached or read. Typical questions were: Have you read any of these books yourself? Has someone read any of these books to you? Which of the books do you like the best? Why do you like that book? Do you have any favorite books that you can tell us about? Which book seems the easiest? Which one seems the hardest? Why?

It was noted during these interviews that the children were very accommodating. That is, they obligingly indicated that they thought all of our selections were good books for a library. However, their reasons for selecting books for themselves were sometimes uninformative, and occasionally misleading. Among the reasons given for preferences were: full color illustrations, big letters, not so much printing, has a lot of imagination in it, has rhymes, is funny, is interesting (often followed by specific comments such as, "I like ghosts," or "turtles are interesting pets"), or simply, "It looks like a nice book," and/or "It's a good book."

There was no apparent consistency across children as to which type of book was liked the best, but humor seemed to be a recurring positive attribute. Books that were rejected by the children tended to be those they thought would be difficult, and difficulty seemed to be judged by print size, size of words and the amount of print in relation to the number of illustrations.

After these interviews, a second procedure for gathering impressions of the children's interest in the books was devised. Children six to nine years old attending an activity-centered summer session of a Boston City School were seen individually by the researchers several times in an effort to have them read books of varying difficulty. During these sessions, the children read to the same researcher, gave opinions about the books, and answered questions on their content. Information was also

kept on which books the children elected to continue reading after the interview.

Eventually, 80 books from the 170 purchased were selected to be included in the study. Books were eliminated on the basis of the children's reactions as judged by the researchers who used the books with the children. Among the characteristics of the eliminated books, several were quite apparent: unusual or difficult format or typography; illustrations that were not liked; repetitious content; unusual language (such as foreign expressions, difficult verse, or unnatural language patterns); abstract concepts and those foreign to the child's own experience; stories which began at a very slow pace or very tediously; watered-down versions of classics; those with elements of moral preaching; and those which were of interest to much younger children and written as though they were to be read to children by adults.

Our procedures for the selection of books most certainly resulted in some idiosyncratic decisions. Other researchers would have had their own notions about including certain books and rejecting others, and since our selection has not been validated in any way, we do not think it is fair to report the titles of books which we eventually eliminated.

The final corpus of 80 books selected is given in Appendix A. The procedures used yielded a set of books that is representative of the total universe of trade books children might like to read. While most books selected were very popular with youngsters during the time of the study, they might not be at another time or in another place. The description of our book selection procedures has been given to permit the reader to determine for himself how representative our book collection may be of the universe of books one might be interested in using as the basis of a readability measure for children's trade books.

III. MATCHING CHILDREN TO THE BOOKS

The Problem

Given the corpus of 80 books and a population of second and third grade children from a middle to lower SES public school in the City of Somerville, Mass. who were to provide difficulty data by reading the books, random matching of books and children would have been an inefficient process because books that were much too difficult for a child would have produced nothing but reading errors and books that were too easy would have produced no error data at all. Furthermore, the frustration and boredom of children reading excessively difficult or simple books might well have introduced unsystematic factors into the difficulty data. In order to avoid these problems, each child was matched up with a group of books appropriate to his level of reading skill. In this way, each child produced reading error data that served to calibrate the difficulty of one or more groups of books.

In order to match children and books, it was necessary to arrange the books in an approximate order of difficulty and to make use of some procedure for assessing the children's reading skills with respect to the books. After considering a number of alternative procedures, such as trial and error or the use of readability measures and standardized reading tests, it was decided that the most direct and probably most valid procedures lay within the corpus of books themselves. The strategy adopted was to rank order the books by judgment, using the concensus method, then to construct placement tests made up of systematic, rank-order difficulty samples of the scaled books. The initial task was to provide an approximate ranking of the books.

Ranking the books

In the beginning, there was little confusion, for it was clear that the books varied in difficulty as well as in other features such as physical size, length, style, illustration and so forth. The task was to put the books into rank order of difficulty and into equivalent groups of a manageable size. A series of tentative sortings of the books made it clear that one could rank-order sets of ten books without great effort, and that twenty books could be ranked after two or three considerations of the same set. As a result of these experiences, it was decided to assemble the books into sets of 20, each set spanning the range from easiest to most difficult. These sets were the basic materials for the entire study.

Four sets of 20 books each were assembled by the following procedures. Over repeated meetings, four to seven researchers sorted the books, according to their combined judgments, into five groups ranging from most to least difficult. The books within each group were then ranked by the same procedures. These concensus rankings were checked out by listening to children read the books again, and any indicated adjustments in rank order of difficulty were made. Books of similar difficulty were assigned to separate sets until four sets of 20 books each were established; each set arranged in order of difficulty, approximately equivalent to one another in difficulty range, and, insofar as possible, similar in other attributes such as content and interest. These four sets of 20 books each, Sets I, II, III and IV, were the basic sets of books upon which the readability measures were developed and validated. They provided the basis of the placement tests for matching

children and books and criterion performance data was gathered as each child read books from one of the four sets.

Constructing the placement tests

Two logically equivalent cloze placement[1] tests were constructed from Sets I and II of the ranked books. Form A was composed of selections from books at levels 1, 3, 6, 10, 13, 16, and 20 of Set I, and Form B from levels 1, 3, 6, 10, 13, 17, and 20 of Set II. One hundred and fifty word passages were selected for the tests from each book. These selections were taken from the beginning of each book so that understanding would not depend upon previous story lines. In addition, the first twenty to fifty words of each book were presented in tact prior to the cloze test passages, so that children could warm up and become accustomed to the test situation before facing the test passages.

The 150 word ranked passages were converted into cloze test items by application of the following rules:

1. Select every 15th word.

2. If the 15th word is a content word (noun, verb, adverb, or adjective) delete and use as an item.

3. If the 15th word cannot be guessed from the preceding context; if it is a repeat of a word previously deleted several times; or, if it is not a content word, delete the closest word that meets rules 1 and 2.

The above modifications of the "standard" cloze procedure (deleting every 5th item) were adopted after numerous trials with differing deletion ratios. A deletion every fifteenth word preserved enough of the normal reading task to prevent frustration and/or great modification of "real-world" reading behavior.

The resulting items were tried on six adults who were asked to guess the deletions on the basis of preceding text and two words beyond each deletion. Items not guessed correctly by these adults were replaced (following rule 3 above) on the assumption that children would find them too difficult.

Two tests were assembled out of xeroxed pages of the cloze-deleted selections from the books. By using copies of actual book pages, the illustrations, type style, layout, and other "bookmaking" features of the selections, except for color, were preserved, for whatever value they may have contributed to readability or child response to the stories. Each form of the test consisted of seven passages of 150 words each, representing seven different levels of difficulty. An additional cloze reading selection preceded each test as a demonstration and training item. A list of passages making up the tests and training items will be found in Appendix B.

The test passages were presented in an order designed to bracket a range for each child's level of reading skill. The order was also designed so that he would not have all of the easy or difficult passages at one time, and so that the less competent readers would not have to struggle with the most difficult passages.[2]

Administration and scoring of placement tests

Each participating child was administered one form of the Placement Test, whichever form did not contain selections from the set of books he was to read to provide criterion data. That is, Form A was administered to children who were to read books from Set II; Form B was administered to those who were to read Set I books; and both forms were administered randomly to children who were to read Set III books. Each child was tested individually, following the protocol described in Appendix B,

and each child's reading of the test passages was tape-recorded for later scoring.

The tests were scored in three ways:

1. By all oral reading errors;

2. By exact cloze - only the exact deleted word was accepted as correct;

3. By approximate cloze - appropriate synonyms to the exact cloze word were accepted.

These three different forms of scores were investigated to select the most suitable for selecting which books each child would read. This work, described in Section IV, below, resulted in selection of the exact cloze scores.

Using the exact cloze scores, the following procedures were used to match each child with the 10 books he was to read.[3]

1. Using the scores from each level of the placement tests, the mean and standard deviation of these scores were calculated for each child.

2. Each child's own standard deviation was added to his mean score to yield a maximum error score. This maximum error score defined the level of the most difficult book a child would be assigned to read. That is, the highest level passage in the test on which he received this maximum error score was designated as the level of the most difficult book in the set he would be assigned to read to provide difficulty data.

3. Each child's set of books ranged down ten levels from this "most difficult" book.

As a result of these procedures, every child was presented with a set of ten books well within a range of difficulty which he might be

expected to handle without frustration. The better readers thus provided difficulty data on the harder books, the poorer readers on the easier books.

FOOTNOTES

1
The "cloze procedure" is discussed more fully in Section IV, page 28

2
The testing sequence and entire protocol for placement test administration is given in Appendix B.

3
An example of this scoring procedure is given in Appendix C.

IV. DEVELOPMENT OF A CRITERION

Over the years, a number of different criteria have been used in the development and validation of readability formulas. According to Klare's (1963) analysis, comprehension, judgment, speed, readership, listenability, and writer characteristics have all been used, with the first three being the more popular. Beginning with the work or Lorge (1939), the most popular comprehension criterion probably has been the 376 reading selections in the McCall-Crabbs Standard Test Lessons in Reading (1926), which was standardized according to the number of correct responses to comprehension questions on each selection. These same passages also were used by Flesch (1943) and by Dale and Chall (1948) in devising their formulas. Later, Farr, Jenkins, and Paterson (1951), Forbes and Cottle (1953), and Fry (1968) used the Dale-Chall and/or Flesch formulas to cross-validate their own work. Cross-validation has been a popular procedure in the development of readability formulas, as in other areas of measurement, probably because of the difficulty of developing an adequate primary criterion.

The grade level structure of schools and school books has led to the use of assigned grade level as a criterion in developing formulas to be used on reading materials for younger children (grade three and below). Grade levels assigned by publishers to books in the basal reading series were used as a criterion variable by Dolch (1928), DeLong (1938), Stone (1938), Dolch (1948), Wheeler and Smith (1954), and Bloomer (1959). Spache (1953) and Johnson (1930) used the grade level of classrooms which were using the primary textbooks under study. Washburne and

Morphett (1935) assigned grade levels (1-9) to children's books and used these as their criterion variable. The grade level they assigned to each book was the median grade level score obtained on the paragraph meaning section of the Stanford Achievement Test for those children who had read and liked that specific children's book. In an earlier study (Vogel and Washburne, 1928), seven-hundred books had been scaled (each one having been read by at least 25 of the 37,000 children tested in grades 3-9), and 152 of these were used in constructing their formula.

More recently, Bormuth (1964) constructed cloze tests on twenty 275 word passages in the areas of history, geography, biological science and physical science, varying in level of difficulty from about fourth grade to eighth grade. His criterion measure of passage comprehension difficulty was the mean of "word difficulties" in the passage, found by determining the proportion of 139 subjects passing an item (word) on the cloze test.

There are clear choices to be made among the above criteria, both for the development and the validation of readability measures. Use of a primary criterion is preferable to a secondary criterion such as provided by cross-validation against an existing measure; a criterion that samples behaviors such as the reader's comprehension, speed, or his choice of reading matter is preferable to one that uses judgments by individuals not in the target population of readers. Although the grade-level criteria cited above may be suitable for developing readability measures for school books, there is no reason to assume they would be valid for children's trade books, which are not always written with a grade level target in mind. Among the above criteria, the McCall-Crabbs Standard Test Lessons come closest to providing a suitable

primary criterion; they are scaled, they sample the behavioral repertoire of "comprehension," and have a known relationship to existing readability formulas. Unfortunately, they were unsuitable for our research on several counts: the text of the McCall-Crabbs is very much like school-book texts and very little like the prose of children's trade books; the difficulty range covered is too high for lower elementary children; and, the form and character of the multiple choice items used adds unspecified difficulty factors to those of the text.

Such considerations led to our decision to develop a criterion specific to our own purposes. Since the purpose of our readability measure is to assign to trade books scores which reflect the difficulty of the books for beginning readers, we elected to use as a criterion the data created by children actually reading a sample of trade books. These data were to be gathered from an independent sample of the universe of trade books for children (see Part II). One set of data was to be used in the formulation of the measure, and another set in validating the measure. The question which remained was: What aspects of reading behavior should be represented on the criterion?

Reading is a complex behavioral process with a large universe of behaviors that could, in theory, be used as readability criterion. A person may read silently or aloud, follow directions, search for information, answer questions based upon the reading, become happy or sad, read with or without vocal and facial expression, and so on. Reading is all of these behaviors and others not named; the nature of the criterion depends upon which of these behaviors is sampled, and in what manner. An ideal readability measure might be one that would predict which books a child would select, read, understand, enjoy, and

finish; the criterion for developing such a measure would have to sample each of those behaviors and relate them to textual variables. Such an undertaking was clearly utopian, so we have followed a modest course, one more in keeping with the traditions of past readability research, and developed a criterion of difficulty based upon comprehension. Comprehension is, in fact, the basic purpose of reading, it is the behavior which reinforces the reader and prompts him to continue reading. One may take some comfort in the belief that desirable reading behaviors such as enjoyment and completion of a book will not exist for readers with too difficult a book, but may exist with a book of suitable difficulty.

The measurement of comprehension is encumbered by theoretical and operational problems as complex as those facing the measurement of intelligence, suggesting that one may define comprehension as "whatever is measured by comprehension tests." A recent study by Auerbach (1971) shows that standardized comprehension tests are made up of a large array of items, tapping a variety of behaviors, assembled quite unsystematically into tests. The resulting tests cannot be said to provide a systematic sample of the universe of behaviors called "comprehension," especially since the ultimate composition of the test is influenced by a process of item analysis which selects discriminating items rather than representative items. Nor is comprehension a clear theoretical construct. As developed in the context of schooling, comprehension has tended to mean the retention of "content," as measured by multiple-choice tests administered sometime after reading has taken place. Comprehension can equally well refer to the "concurrent" cognitive and affective understanding of a novel or following directions while building something, neither of which have substantial retention components. It was felt

that a comprehension measure which minimizes the retention factor was desirable as a difficulty criterion for trade books, since they are generally read for pleasure and incidental learning, not for assigned acquisition of knowledge. With this rationale in mind, a choice needed to be made between reading rate, oral reading errors, multiple choice items, or cloze items.

Reading Rate

Data on the reading rate, or speed at which the children read the various selections is perhaps a viable criterion, but it is an indirect measure of comprehension. The factors influencing rate of reading are many because as the reading selection varies, the task itself varies. Better comprehension could well be evidenced by slower reading for some types of material.

Oral Reading Errors

Errors made by a child in oral reading have strong face validity as a criterion measure of reading comprehension. More of the process is available for observation than with silent reading, where selected observations must be made in the form of answers to questions. Intuitively, one assumes that oral reading should provide a valid criterion of comprehension. In all probability it could reflect directly what the child does when he reads, but there it is not clear whether or not child understands what he is able to read aloud. Similarly, the misreading of certain words in a selection is not always an indication that the child does not understand. Oral errors, such as substituting words for the ones written, mispronouncing proper names, deleting words, and/or inserting words, do not necessarily reflect misunderstanding. However, since the correlation of oral reading errors and comprehension scores on standard-

ized reading tests have been high for grades 1, 2 and 3 (Bond and Dykstra, 1967), we decided to gather data on the children's oral reading of the books as one means of assessing the validity of the comprehension criterion finally selected.

Although earlier experience in the schools had convinced us that certain types of oral reading errors have little influence upon children's comprehension, nevertheless, we decided to use all errors as a measure of oral reading. This decision was based upon the grounds that we were interested in scaling books, not children, and that any idiosyncratic oral reading responses of a given child would carry through all of that child's reading and therefore equally influence his reading of all the books. That is, we hypothesized that if a child tends to repeat words frequently during oral reading, his/her oral reading score would be reduced across all books, but different children's scores would be averaged for each book, so the influence of any one type of error attributed to a particular child would influence the scores across all the books which he read. And in fact, one might assume that such idiosyncratic errors would become more prevalent as the reading became more difficult and to disregard them in the oral error scoring would throw away relevant data.

In order to substantiate this line of reasoning we had a group of children, not from the population to be tested, read selections from books on level 5 through 18 and scored their oral errors in two ways: (1) counting all errors transcribed according to the scheme in Figure 1 and (2) counting only those errors from Figure 1 which were assumed to indicate misunderstanding of the text or would lead to misunderstanding. Thus no particular type of error was consistently counted or not counted

in the latter method of scoring. The results of that comparison indicated that the relative difficulty of the books assessed by either method remained substantially the same (Figure 2). In addition, scoring all errors should prove more objective and reliable since no judgments have to be made about whether or not an error does or does not fit the context. Errors that were assumed to be the result of a speech problem or a different idiolect or dialect were not counted and a scheme for scoring whole line or paragraph omissions was adopted (see "Rules" below) so as to avoid fallaciously high error scores.

- - - - - - - - - - - - - - - - - -
Insert Figures 1 and 2 about here
- - - - - - - - - - - - - - - - - -

Our scoring for oral errors was based on these rational decisions, but it should be noted again that this scheme is perhaps more suitable for averaging different children's scores in order to rank order the difficulty of books, as was our purpose, than for judging individual children's reading skills.

Rules for scoring oral reading errors. The final set of rules which was used to transcribe and score each child's reading of each book selection follows:

Record all errors on the appropriate sheet (a mimeographed copy of the text) using symbols from Figure 1.

Reproduce as exactly as possible what the child actually said.

Scoring
(Each word can be scored for no more than one of the first six types of errors)

1. Prompts - each word prompted counts as one error
2. Sounding out - each word sounded out counts as one error

NOTATION

prompt after approximately 10 seconds =	p
substitution =	(write in the spoken word)
omission =	(write in word omitted)
word addition or insertion =	the Ad man
reverse order =	
incorrect pronunciation =	
repetition =	
sounding out of the word by phonics or other method =	(underline)
unclassified error =	x

FIG. 1. Symbols used for recording oral reading errors.

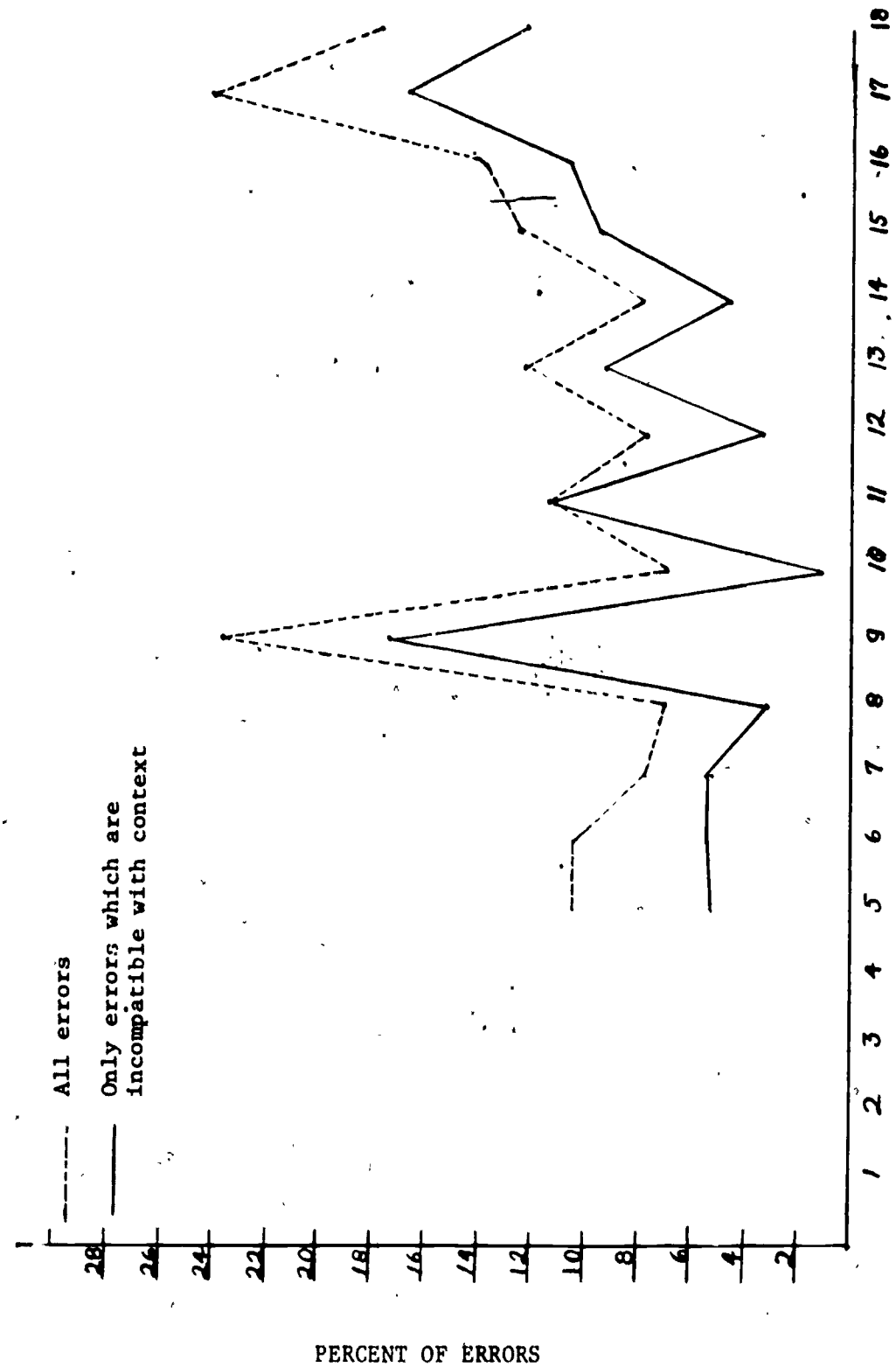


FIG. 2. Comparison of two schemes for counting oral errors.

3. Substitutions - each word substituted counts as one error.
4. Reversals - each reversal of two words counts as one error
5. Word Addition - each word added is counted as one error
6. Unclassified Error - each unclassified error counts as one error
7. Omissions -
   a. within a line, every word omitted is counted as an error
   b. when a whole line and/or sentence or more is omitted, count as one error per line
   c. when a page or more is omitted count as one error per line
   d. when S goes back and reads previously omitted material, retain the errors counted for the omission and count all other errors made in the usual way
8. Repetitions -
   a. within a line, every word repeated is counted as an error; any other reading errors that occur on those same words are also counted
   b. when a whole line or more is repeated count as one error per line
   c. when a page or more is repeated count as one error per line
   d. when material is reread, all reading errors are counted for each reading as often as they occur (except for mispronunciations noted below)
9. Mispronunciations -
   a. a mispronounced word is counted as an error the first three times the same word is mispronounced in the reading passage at the rate of one error for each mispronunciation
   b. if the mispronunciation seems to be the result of dialect or speech problems, do not count any errors, but do mark the words that are mispronounced
   c. if the same word has been mispronounced at least three times and is later given a correct pronunciation during the reading of the passage; subtract one of the three errors (accumulated in 9a above). Thus leaving only two errors for the repeated mispronunciations.

Multiple Choice Test Items

A comprehension measure where a child has to act on what he has read, make a decision, or perform a specific act has certain advantages

over both speed and oral error measures. Multiple choice type test items containing a variety of questions falling into categories such as main ideas, facts (or details), and inferences is one such measure that has gained widespread use. However, it is difficult to create multiple choice test items which are consistent with the difficulty and other characteristics of the passage. Elley's research (1967) also cautions us to give serious consieration to the possibility that questions are answerable without reference to the reading passages at all, i.e., they question facts already known to the reader.

Multiple choice tests contain alternative choice answers which are usually taken directly from the text or else are some sort of transformation of the language in the text. In the former case, a correct response may be given with no comprehension at all; it may simply be a task of locating those words in the text. In the latter case, the child must be able to make linguistic transformations to complete the task, but these would not necessarily be transformations that would be required for comprehension of the passages. Additionally, it it most difficult, if not impossible, to adequately sample all of the content in the passage in constructing multiple choice items. These disadvantages of multiple choice tests led to our rejection of them as a criterion measure for our study.

Cloze Tests

We turned, instead, to another measure of comprehension noted earlier in our description of placement tests, the "cloze procedure." First introduced by Wilson Taylor (1953, 1956), it became prominent in the 1960's. In cloze tests, a certain number of words from the reading selection (most often every n^{th} word), are deleted and

replaced with a standard-size blank. Subjects are asked to fill in the missing words as they read the selection. The "cloze score" for a passage is equal to the number of correctly guessed missing words. Taylor's 1956 study made use of cloze procedure in assessing readability:

> It was found that cloze scores repeatedly ranked three "standard" passages in the same way the Flesch "reading ease" and the Dale-Chall formulas did; and this finding held for four different mutilation systems-one counted out every fifth word, another every seventh, another every tenth, and still another took out 10% at random - each deleted an almost entirely different set of specific words.
>
> The cloze method appeared superior to both the Dale-Chall and Flesch formulas for gauging the difficulty of "non-standard" passages. Chosen a priori were an "extremely easy" passage by Erskine Caldwell, a "very hard" one by Gertrude Stein and an "extra hard" one by James Joyce. These passages were included in a set of eight passages used in the experiment. The Stein passage (written in short and familiar words and short sentences) was, although it made very little "sense," scored as "easiest" by both of the formulas. Also, the Flesch formula indicated that the Joyce and Caldwell passages were both "fairly easy." Two different sets of cloze scores, however, agreed in scoring these non-standard passages according to a priori expectations. (p.44)

The cloze procedure has been used as a test of comprehension for individuals as well as for readability measurement. High correlations between cloze tests and standardized reading achievement tests (usually multiple-choice) and/or correlations between cloze tests and readability measures (regression equations) are reported by Jenkinson (1957), Ruddell (1963, 1965), Gallant (1965) and Bormuth (1967).

Selection of the cloze procedure over other comprehension measures as a criterion for our study seemed advisable from both a practical and logical standpoint. Some of the problems inherent in the multiple-choice measures could be overcome by the cloze technique. In a cloze test, the language of the passage itself is the only

language used; the difficulty of the cloze item is a function of the way the passage is written, rather than the way in which a question is written. A modified cloze procedure which offers alternative words for the reader to select for the blanks was not used because such items confirm that one of the given alternatives is correct. This modification is subject to many of the criticisms of the multiple choice tests noted above.

The task for the student in completing a cloze passage is to respond to all of the text in reconstructing an appropriate word at every deletion. Because words are deleted "in situ," responses to the regular cloze procedure involve a simultaneous understanding of the syntax and the semantics used by the author. We hypothesized that cloze passages are less likely to destroy the student's typical reading behavior, that his performance on cloze tests is closer to what he does when he is reading outside a test situation.

For the above reasons, the cloze procedure was selected as a criterion measure for scaling the sets of books. Essentially, we agreed with Potter (1968) that the cloze procedure was a "method for intercepting the message from the transmitter or the author, by mutilating its language pattern and administering it to receivers or readers in such a way that their attempts to make the patterns whole again will potentially yield a measure of their ability to deal with the general meaning and form of the passage." Variables which might affect the scores of children tested using a cloze procedure seemed more self-evident and more contained within the reading selection (and the reader) than the myriad of additional variables which underlie tests whose itemsgo beyond the passages to be read.

We did not wish to constrain our early readers' responses to the cloze items by requiring them to be written. As we had already decided to have the passages read orally, we chose to record each child's response to each item.

Scoring Cloze Passages

In scoring cloze passages, a decision had to be made about what constituted a correct response to each item: the "exact word" used by the author, or any word that could be considered a semantically and grammatically correct "synonym" which made sense in the context of the selection. Taylor's (1953) study compared scoring "precise matches only" (exact word) with "matches plus synonyms," and allowing 1/2 count for each synonym, he concluded that the degree of differentiation between the books he was testing was "virtually identical" (p. 425) under both schemes. Other researchers, Rankin (1957) and Ruddell (1963) found only slightly increased variances on individuals' scores when they included synonyms. We decided to score our placement tests both ways and make a comparison between the two cloze scoring methods, the oral error data[1] (see Figures 3 and 4), and our earlier preliminary rankings of the books (see page 12, Section III). Both cloze scoring methods gave a good range of differences across all seven selections. Inspection of these data suggest that the exact word scoring gave a steeper slope and therefore would be preferable for discriminating between books. (See Figures 5 and 6.)

- - - - - - - - - - - - - - - - - - -
Insert Figures 3-6 about here
- - - - - - - - - - - - - - - - - - -

Rank order correlations were computed between three scoring methods: oral, exact word, and synonyms (Tables 1 and 2).

- - - - - - - - - - - - - - - - - - -
Insert Tables 1 and 2 about here
- - - - - - - - - - - - - - - - - - -

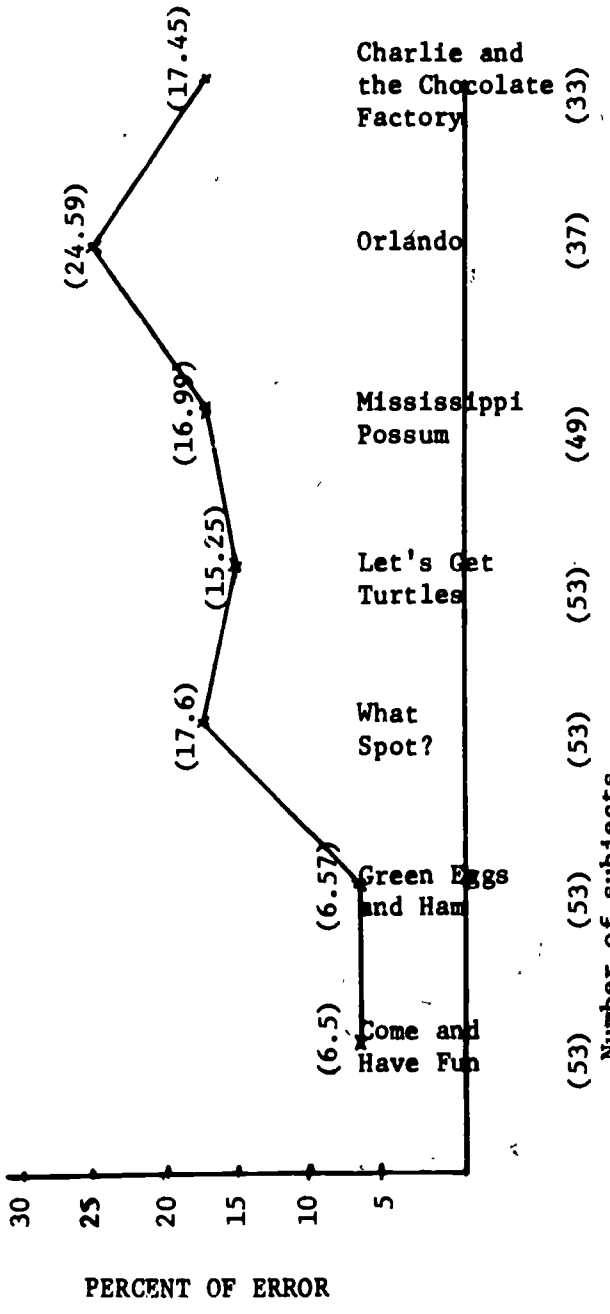


FIG. 3. Mean percent of oral errors on seven selections in placement test (Form A)

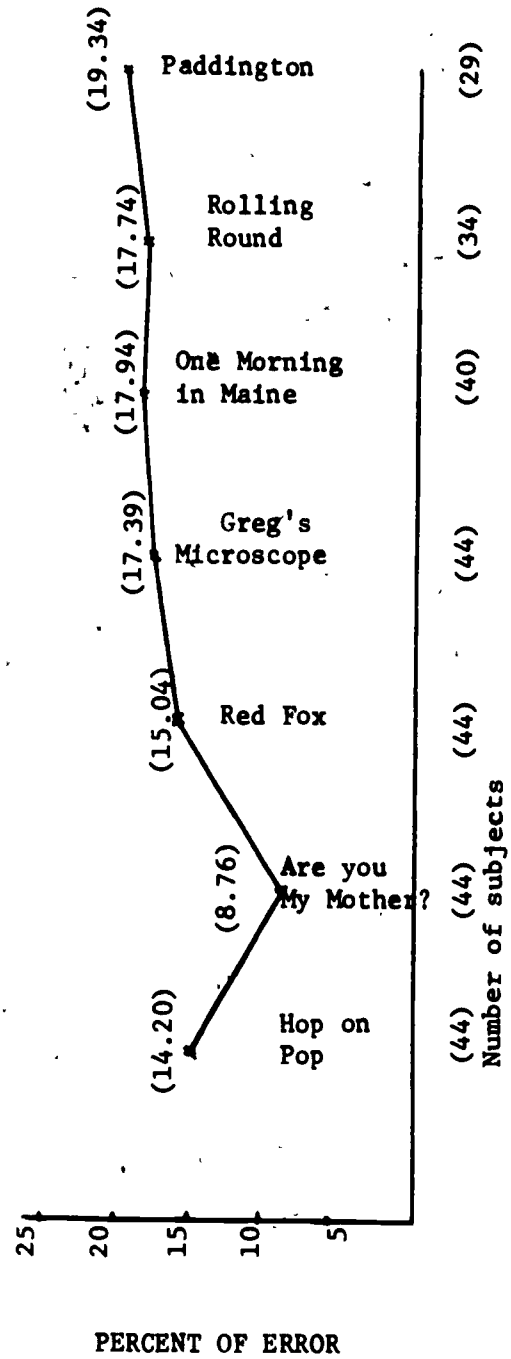


FIG. 4. Mean percent of oral errors on seven selections in placement test (Form B)

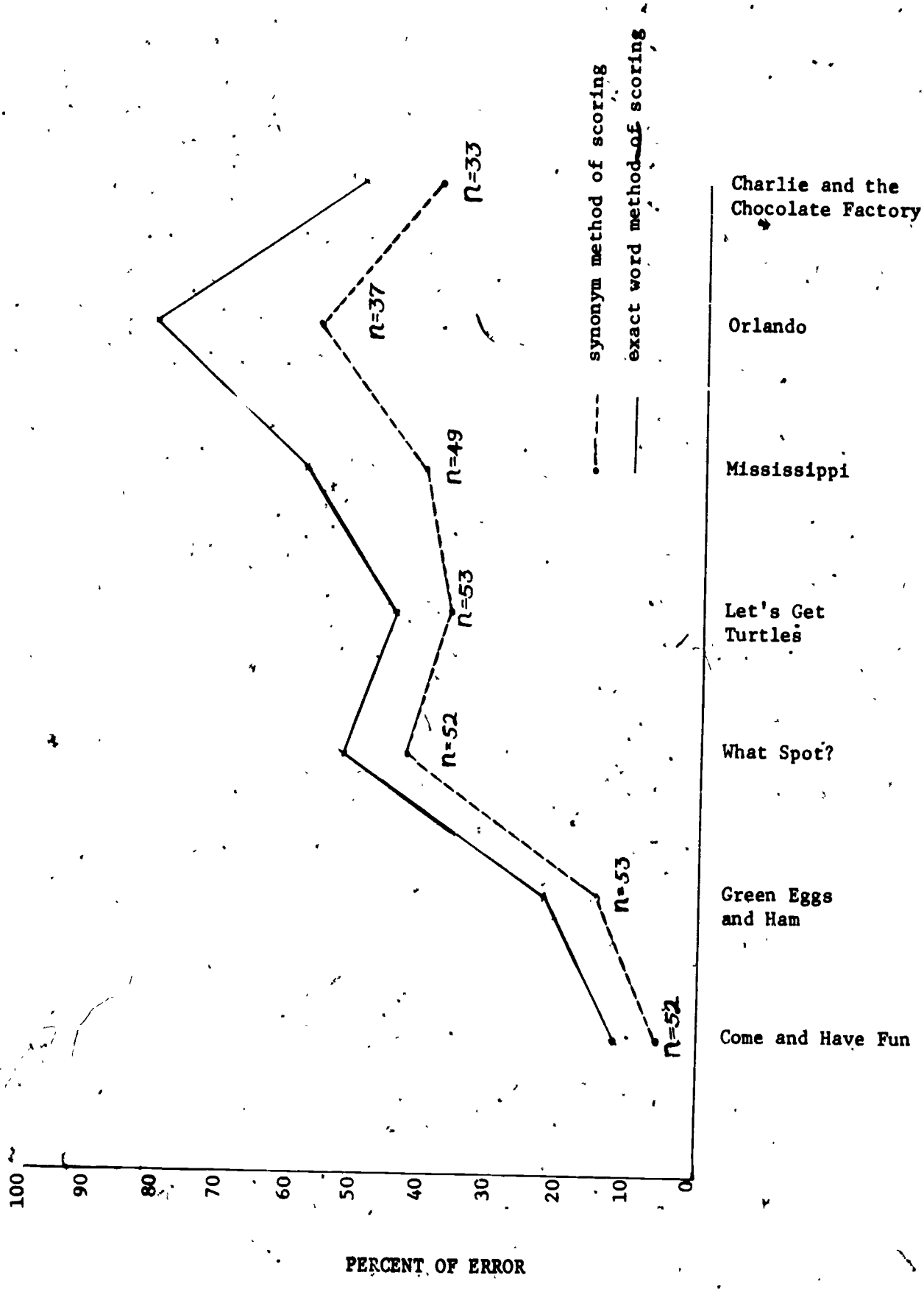


FIG. 5. Mean percent of cloze errors on placement test (Form A): Two methods of scoring.

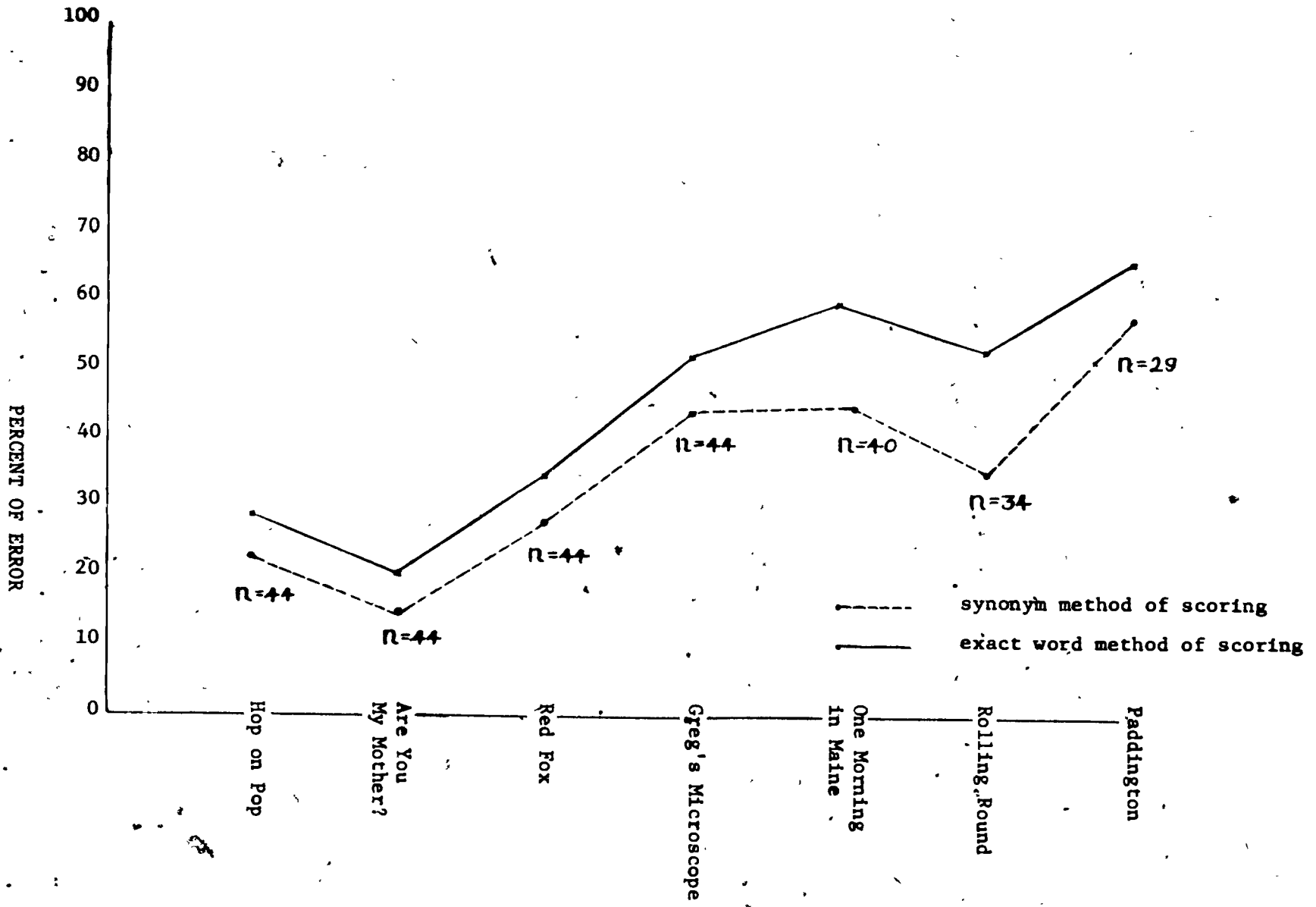


FIG. 6. Mean percent of cloze errors on placement test (Form B): Two methods of scoring.

TABLE 1

Intercorrelation Matrix for Rank Order Correlations* Between Various Scoring Methods on the Seven Selections in Placement Test Form A

	Oral reading errors	Exact word cloze	Synonym cloze	Researcher assigned provisional ranks
Oral reading errors	--	.86	.89	.86
Exact word cloze	.86	--	.96	.71
Synonym cloze	.89	.96	--	.65
Researcher assigned ranks	.86	.71	.65	--

TABLE 2

Intercorrelation Matrix for Rank Order Correlations* Between Various Scoring Methods on the Seven Selections in Placement Test Form B

	Oral reading errors	Exact word cloze	Synonym cloze	Researcher assigned provisional ranks
Oral reading errors	--	1.00	.89	.93
Exact word cloze	1.00	--	.89	.93
Synonym cloze	.89	.89	--	.82
Researcher assigned ranks	.93	.93	.82	--

* $R' = 1 - \frac{6\ d^2}{N(N^2-1)}$

The generally high correlations among the measures indicate that (with books spaced on the difficulty contiuum such as those used in the placement tests) all three measures are suitable for rank ordering the difficulty of the books. The exact word method and the synonym method of scoring the cloze items are highly correlated. The exact word method has slightly higher correlations with the oral scoring and the researcher assigned provisional ranks than does the synonym method in three of the four comparisons.

In addition to providing somewhat greater differentiation among books and correlating with oral scoring as well as or somewhat better than the synonym method, the final consideration that tipped the balance in favor of the exact word method is the fact that it is an easier and more objective method of scoring (no other word than the one used in the text is acceptable). We therefore decided to score the cloze passages using the exact word method as our criterion variable and to continue to collect oral error data as a check on the validity of the exact word measure.

FOOTNOTES

[1]Oral reading error scores consisted of the number of errors made divided by the number of words in the selection (see Figures 3 and 4). The cloze items (i.e., the words deleted) were not included as either "errors" or "words in the selection" in the computation.

V. CRITERION DATA COLLECTION PROCEDURES

Subjects.

Criterion data on difficulty of the books was obtained from children enrolled in the second and third grades of a neighborhood school in a low socioeconomic section of a city of 88,000 adjacent to Boston. Children were selected at random, and only the few children who were not considered "independent readers" by their teachers did not participate. Ultimately, 197 children participated by reading those books appropriate to their levels of skill. All had been taught reading through a structured basal reading method based primarily on the sight approach, with an additional phonic component in the intital stages.

Measures.

The Placement Test, either Form A or Form B, was given to each child. Cloze passages in each of 60 books, 20 in each of Sets I, II, and III, were prepared in the following manner.

Passages of 150-175 words were selected to be representative of the general tone and content of each book, taken from as near the beginning of the book as possible. In most cases, a few pages preceded the cloze selection to give the children an opportunity to read orally and familiarize themselves with a book before attempting the cloze selection.

The cloze passages were prepared by deleting every 8th word, a compromise between every fifth word, used by many researchers (Taylor, 1953), and a much lighter deletion ratio which would have yielded too

little information. Pilot work had shown that a ratio of 1:5 was disruptive of children's usual reading patterns, but that 1:8 did not remove enough context to interfere very much with reading even when the deleted word could not be guessed.

In his original presentation on the cloze procedure, Taylor (1953, p. 420) defends deletion of every n^{th} word regardless of the "importance" or the grammatical form-class of the word. Intuitively, however, one can hypothesize that the meaning or information load, the substantive content of written text, resides more in the descriptors and operators -- the nouns, verbs, adjectives and adverbs, than in the function words. Since Rankin's study (1959) tended to support such a hypothesis, we decided to delete all form classes except prepositions and conjunctions. The complete set of rules followed for deleting words to construct cloze passages are given in Appendix D, page 114 .

The cloze passages were constructed in the books by masking out the chosen words with pieces of opaque adhesive tape. The procedures followed for a conventional cloze technique (i.e., replacing the deleted words with a standard-size blank (Taylor, 1953) did not seem appropriate for our purposes as we wanted the children to be reading in the actual books with all of their variable formats, thus preserving all the influences of book illustrations, size and style of type, and other characteristics. Any additional prompt to guessing the deleted word which might come from observing its comparative length, we argued, would be a close approximation to an actual reading situation, and would not affect relative book difficulty since it would be uniform across all books.

These same cloze passages were also to be scored for oral errors

as supplementary data. Of course, the responses given for the deleted words would never be scored as oral errors.

Procedures.

Children were taken, one at a time, to a quiet room in the school building where a tape recorder was set up. The Placement Test was administered during the first session, with most children completing the Test and returning to their classrooms within 20 minutes. The protocol followed for placement testing will be found in Appendix B, page 107.

After determining from the Placement Test performance which levels of books a given child would read (see page 15), he was taken on subsequent days to the same room and asked to read the prepared cloze passages in the appropriate books. Some children's scores on the Placement Test indicated that data should be collected from only the six easiest books and therefore not every child read ten books. A few children completed all readings in one session and a few took as many as five sessions. Each was reminded of the cloze task and the same procedures were followed as outlined in Appendix B for administration of the Placement Test (tape recording, moving to a position slightly behind the student, not prompting, periodic reinforcement, etc.).

Scoring.

The selections read by each child were scored by first listening to all tapes and marking errors on a typed facsimile of the text of each selection. Oral errors were tabulated (according to procedures described on page 24) and the number of oral errors was divided by the number of words in the selection, giving a percentage score for oral errors.

The responses to the cloze items were transcribed and the number of

nonexact word matches was divided by the number of cloze items, giving a percentage score for cloze errors.

Not all books were read by all children. The less able readers read only from the easier books beginning with our provisional level one, others read a set of ten books beginning with level 2 or 3 and so on. A scheme had to be developed to adjust the data so that scores would be comparable across all 20 books; i.e., the data needed to be the facsimile of scores of a group of children who had read all 20 books. The following procedure was adopted. Scores from all the children reading books on provisional levels 1-6 were grouped together and a mean error score for each of the books was calculated. Similarly, mean error scores for books on provisional levels 3-8 were calculated from error scores of the children who read these books. The mean difference between the two group means was calculated for each of the four overlapping books (3,4,5 and 6). This mean difference for the overlapping books (-12.12 in the example for Set III illustrated in Figure 7) was then added to the scores of books 3-8, thus adjusting this set of books to the base line of the lowest readers. (See Figure 7.) The same procedure was followed to adjust the scores of each set of books up through the 20 levels.

- - - - - - - - - - - - - - -
Insert Figure 7 about here
- - - - - - - - - - - - - - -

In order to carry out this scheme of adjusting scores, the percentage error scores were tabulated into matrices of "children by books." Subsets of books for computing adjusted scores were formed by grouping together those books that had been read by at least six children and then grouping another more difficulty set that would both overlap the first set and extend beyond it. This procedure was followed

Book Level	Mean % error, first group of Ss	Mean % error, second group of Ss	Difference between mean scores	Mean differences for four overlapping books	Adjusted cloze error score on books 3-8 (actual score plus "mean difference" of 4 overlapping books)
1	36.00				
2	32.50				
3	47.50	29.52	-17.98		41.64
4	55.50	50.00	-5.50	-12.12	62.12
5	56.00	45.00	-11.00		57.12
6	58.54	44.56	-13.98		56.68
7		63.00			74.12
8		61.00			73.12

FIG. 7. Example of adjusting cloze error scores to the baseline of the lowest readers.

for Sets I, II, and III for both oral error scores and cloze scores.

Results. The adjusted cloze scores and oral error scores are presented in Tables 3, 4, and 5 for Sets I, II, and III. It can be seen that the resulting adjusted scores can no longer be designated as percentages of error-scores although they are based upon percent of error scores. This is because the percentages have been adjusted to the base line of the lowest readers reading the easiest books.

Insert Tables 3, 4, and 5 about here

The results of the above empirical scaling were correlated with the researcher assigned provisional rankings for each set of books. It is clear that the relationships between the cloze scores, the oral error scores and the researcher assigned provisional rankings are very high. These correlations, given in Table 6, are all significant at the $p < .01$ level.

Insert Table 6 about here

Methodological Problems.

Data were collected and analyzed as described above for three sets of 20 books each. These data were comparable only within each set of 20 because children did not read across sets. Any given group of children read books in one set only and thus there was no evidence that

TABLE 3

Adjusted Cloze Error and Oral Error Scores for Books in Set I

Books	Cloze error score	Oral error score
Come and Have Fun	52.2	18.71
Who Will be My Friends	50.6	17.12
Green Eggs and Ham	27.3	13.29
Summer	57.5	20.09
Little Bear's Visit	87.4	19.56
What Spot?	65.8	23.84
Case of the Hungry Stranger	75.9	27.49
Shhhh.....Bang	78.4	28.83
Here Comes the Strikeout	73.7	25.56
Let's Get Turtles	84.2	29.04
Blueberries for Sal	71.9	29.58
Mike Mulligan	89.5	29.31
Mississippi Possum	78.7	31.74
Where Does Everyone Go?	100.5	33.30
Yertle the Turtle	99.6	35.42
Orlando, the Brave Vulture	89.8	37.06
Camel in the Sea	92.0	35.52
The House on E. 88th Street	98.4	37.18
Anatole and the Robot	108.0	39.40
Charlie and the Chocolate Factory	105.4	35.32

TABLE 4

Adjusted Cloze Error and Oral Error Scores
for Books in Set II

Books	Cloze error score	Oral error score
Hop on Pop	41.5	18.52
Where is Everybody?	66.5	16.85
Are You My Mother?	31.3	13.04
The Bike Lesson	64.1	23.11
I Should Have Stayed in Bed!	62.9	25.76
Red Fox and His Canoe	64.9	24.61
The Case of the Cat's Meow	83.6	32.22
Whistle for Willie	83.7	32.22
Just Me	77.1	27.67
Greg's Microscope	80.1	31.23
Make Way for Ducklings	83.7	30.00
White Snow, Bright Snow	92.6	23.69
One Morning in Maine	61.1	28.81
My Father's Dragon	80.1	33.13
Tico and the Golden Wings	87.6	31.39
John J. Plenty and Fiddler Dan	103.3	32.78
Rolling Round	95.0	36.80
Baba Yaga	104.3	35.52
The Adventures of Beetlekin	105.3	31.64
A Bear Called Paddington	90.1	34.80

TABLE 5

Adjusted Cloze Error and Oral Error Scores
for Books in Set III

Books	Cloze error score	Oral error score
Ten Apples up on Top	36.0	9.96
Put Me in the Zoo	32.5	18.66
King, the Mice & the Cheese	41.6	19.51
Little Bear	62.1	23.42
Shoes for Angela	59.9	23.62
Oliver	69.2	23.49
Snowy Day	67.3	33.06
The Cat in the Hat	70.1	27.10
Bedtime for Francis	75.5	26.82
Fox in Socks	86.0	34.04
Popcorn Dragon	86.9	25.03
Keep Your Mouth Closed Dear	92.5	27.26
Maleline's Rescue	95.3	31.58
Charlotte's Web	111.3	32.03
Zoo, Where Are You?	83.6	29.68
Sam, Bang, and Moonshine	98.6	32.20
Where the Wild Things Are	102.3	31.02
Baron Brandy's Boots	96.9	32.26
The Cookie Tree	91.8	36.11
The Alligator Case	104.8	33.66

TABLE 6

Pearsonian Correlations between Cloze Scores, Oral Error Scores and Researcher Assigned Provisional Rankings for Three Sets of 20 Books Each

	Books in Set I		Books in Set II		Books in Set III	
	Oral Error Score	Rsch. Rank	Oral Error Score	Rsch. Rank	Oral Error Score	Rsch. Rank
Exact cloze score	.890	.864	.820	.818	.838	.910
Oral error score	---	.953	---	.815	---	.825

N = 20

$p < .01$, $r = .492$

comparable data, not only on the 60 books read, but on more books. Klare (1963) offers a useful analysis of three kinds of validity in the evaluation of readability measures. The first is original criterion prediction, "...the extent to which formula scores are related to, or predict, the original criterion scores used in developing the formula." (p. 111); the second is comparative validity data, "...the extent to which scores derived from or more formulas agree with each other." (p. 111); and, the third is validation against outside criteria and it "...concerns the ability of formula scores to predict an 'outside criterion' of readability." (p. 121). Klare suggests that the latter form of validity usually seeks to establish the relationship between formula scores and estimates of readability arrived at in some other way -- comprehension scores, judgments, readership, etc. An alternative and more generalizable method of establishing the form of validity is through the relationship between derived readability formula scores and estimates of readability based upon a different sample of reading material.

From our data, the first form of validity could be determined by deriving a readability formula from a set of books and then comparing the difficulty predicted by the formula with the measured difficulty of the same set of books. Clearly, this form of validity has limited generality since the same set of reading matter is used for both derivation and validation of the readability measure.

The second form of validity, comparative validity, could be determined in the traditional manner by assessing the extent to which our formula correlated with the predictions of other formulas such as the Spache (1953).[2] Originally we undertook this study because no

the sets were equivalent samples of the same universe of young children's literature; in fact there is reason to believe they are not.[1] Equivalence between sets could have been established by having each of the 197 children read sufficient books from the two other sets he had not yet sampled. This was not done because of the large amount of time required (197 children reading 20 selections each yields about 600 hours of data collection). This practical problem led to a search for some other way of establishing comparability across the sets of books.

Despite generally high correlations between the cloze data, oral errors, and researcher assigned provisional ranks, book to book and set to set inconsistencies in the cloze data led us to re-examine its suitability as a criterion for the development of a readability measure.

Variability was introduced into the cloze data by the relatively small sample of each book read by each child, and by each child's prior familiarity with the books, which could not be assessed accurately. Differences between cloze rankings of some books and the researchers' judgments of book difficulty could have been due to the researchers judging difficulty from examination of entire books, where the cloze data were based upon samples of the books.

Had we continued testing until all books in any given set had been read by all children the data would have been more stable, but the practical considerations of time, stated above, prevented this step. The problem faced was how to increase the reliability of the criterion data without having more books read by children. As will be seen from the discussion to follow, (see page 51 ff) the solution adopted was to add more books and to collect comparable data on the entire set.

Validity concerns also prompted us to consider collecting

existing formula had been validated on literature written for young children. An attempt to seek such cross-validation could well be misleading because of differences in the universes of reading matter upon which the criterion data are based.

The third form of validity could be determined from our data by partitioning the books into two sets, using one set for deriving the readability formula, and the other set for validation by predicting its readability. It is clear that this procedure is a more rigorous and useful form of validation because it assesses the generalizability of the derived readability measure to a new set of materials.

These two considerations of increasing the reliability of our criterion data and expanding the set of books sufficiently so that the third form of validity could be determined suggested a revision in the form of criterion data to be used. Since we were unable to obtain further, direct cloze data, a formal judgmental ranking procedure was utilized to expand the total number of books from 60 to 80, and to obtain comparable difficulty data on all 80 books. With this procedure, the original cloze criterion data could serve as a performance criterion once-removed as we could correlate it with the ranking data which would function as the primary criterion.

It was decided to rank all 80 books on one scale of difficulty. Each book would have to be judged as more difficult than all the books ranked below it and easier than all those ranked above it. A subset of 50 of the 80 books (hereafter referred to as the "Formative Set") would then be used in the development of the readability formula and the remaining 30 books (hereafter referred to as the "Validative Set") would be reserved for assessing the validity of the

readability formula that had been developed.

The Formative Set was a modified representative sample of the 80 books. Ten books from each of the three sets of 20 (I, II and III) were selected to be in the Formative Set and the remaining ten from each were assigned to the Validative Set. Selection was made in the following manner: Books in Sets I, II and III were assigned the numbers 1-20 within each set. Then, the first ten instances of the numbers 1 through 20 selected from a table of random numbers designated the ten books from each set to be included in the Formative Set. All 20 of the books in Set IV were assigned to the Formative Set, since no children's cloze scores or oral error data were available on this set, and the books would not be as useful in a validation study. However, it was quite logical to assign this fourth set of books to the formative analysis since judges' rankings of the books would serve as the criterion variable. Hence, the Formative Set includes a random sample of ten books from each of Sets I, II, and III, and all 20 books from Set IV. The Validative Set of 30 books that remained was a random sample of ten books from each of three sets of 20, Sets I, II, and III. A list of the books in the Formative and Validative Sets are included in Appendix A, page 103.

Books were assigned to the Formative and Validative Sets prior to the ranking procedure. However, no indication of this assignment was apparent to the judges during the ranking procedure which dealt with all 80 books as one set.

Rank Ordering of the 80 Books by Judgment

The all-female judges were five undergraduates from Radcliffe College and two graduate students from the Harvard Graduate School of Education. They were selected by interview, after being told about the

requirements of their task, mainly that careful and responsible effort was required. They were clearly above average in intelligence, perseverence, and other personality attributes that would lead to systematic judgmental data, but none had had any specialized training in children's literature.

Training was carried out in one session of 90 minutes, in which the following was accomplished:

1. Purposes of the study were explained.
2. Procedures were described and reviewed.
3. The variables and basis of judgment were explained and practiced.

The variables and basis of judgment were taught as follows. An independent set of ten books was ranked, de novo, by each of the seven judges. The resulting data were displayed, were examined by the group, differences in judgment were discussed, and the bases of judgment were clarified by the experimenters. As a result of this treatment, the judges had common understanding of which characteristics of the books to use in judging difficulty and which should not be used, but no agreement on weighting of the variables. The concensural variables were: vocabulary, language style, structure, complexity, and abstractness. Variables that were not to be used were: interest, type size, book length. Some notions shared by the group follow:

Vocabulary was considered very important, and is judged intuitively by sensing which words would be unfamiliar to a child or difficult to read. As words are repeated over and over in a story, the reading becomes easier. Oftentimes the length of words is a cue to their difficulty.

The structure of the language is a factor of difficulty which can be revealed by longer sentences that include modifying phrases or complex constructions. Repetitions of sentence patterns or phrases within

sentences contribute to increased ease of reading. The poetical devices of rhythm and rhyme can prompt the reader considerably as they impose a given structure and limit the words used in certain instances.

An author's style and the way in which he uses language tends to make a book easier or more difficult. Many figures of speech, considerable fantasy, or allegory, usually make the reading more difficult to understand. Dialogue, on the other hand, often makes reading easier. The complexity of the plot and the story's level of abstractness each greatly modify the other considerations of difficulty.

Pictures often aid the reader to better understand the events in a story as they serve to explain some complex aspect, but they can also hinder the reader's understanding when they do not corroborate the text. Therefore, pictures should be judged always in relation to the story.

The final variable discussed reflected concern that the use of strange type fonts as well as unusual arrangements of print on the page might well be confusing to the children. During earlier trial readings we observed this as children confused the order of the text, oftentimes skipping whole paragraphs, when the print was artistically arranged across several pages in some unusual fashion.

The large number of books to be ranked and the relative scale-closeness of adjacent books required the adoption of ranking procedures that would cut down on the quantity of judgmental work, alleviate the memory load, and focus most judgmental effort upon the close discriminations between books differing only slightly from one another. It is clear that asking the judges to attempt to handle all 80 books as one set would have been aversive because of a gigantic memory overload. The resulting data would have suffered from many forms of unsystematic

behavior perpetrated by frustrated and uncooperative judges. On the other hand, explicit use of the classical procedure of paired comparisons, the underlying model for ranking, would have required each judge to make 3160 comparisons between two books. The unshelving and shelving of the books 6320 times per judge represented an impossible commitment of time.

The procedure finally devised was designed to overcome these problems without compromising data quality. The ranking task was structured so that the judges dealt with sets of books, similar in difficulty, but small enough to prevent memory overload.

The books were tentatively ranked in an approximate order of difficulty by one experimenter and then assigned to 15 overlapping sets of 10 books each. This provided an overlap of five books per set. (see Figure 8). The ten lowest ranking (easiest) were assigned to set 1; the five most difficult books from set 1 plus the next five books on the approximate rank order list were assigned to set 2; the five most difficult books from set 2 plus the next five books on the rank order list were assigned to set 3; and so on until the last assignment to set 15 consisted of the ten most difficult books (tentative ranks 70-80).

- - - - - - - - - - - - - - -
Insert Figure 8 about here
- - - - - - - - - - - - - - -

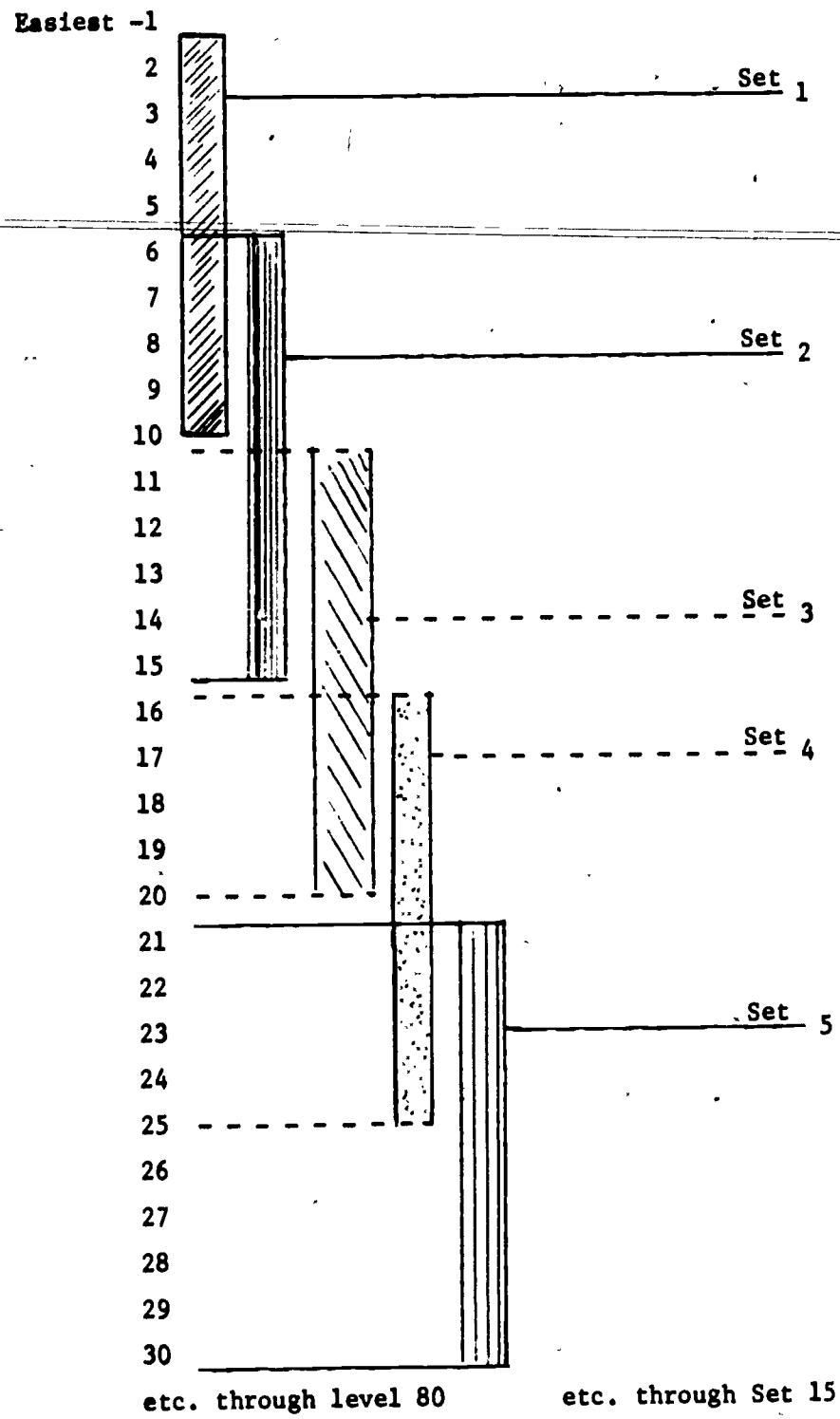


Figure 8. Assignment of individual books to groups of books for ranking procedure

The judges were divided into two groups, one group to rank the sets of books in an ascending order of presumed difficulty, the odd numbered sets first, then even numbered ones. The other group of judges was to rank the sets in a descending order in the same manner, odd numbered ones first, then even numbered ones. (Specific instructions to the judges are given in Appendix E, page 115). By this scheme, each judge was required to rank-order each book two times, once when she ranked the odd numbered sets and again (within a different set) when she ranked the even numbered sets. The two rankings of the same book were separated by intervening judgments of at least six sets.

The books were shelved in the same order throughout the judging procedure and judges removed only the set of books they were working with at any one time. The shelving was based on random numbering of the 80 books and thus the resulting random numbers found in each set of ten books prevented the experimenter's tentative ranking from being revealed to the judges. The judges typically completed the ranking of three or four sets in one session. They dropped their rating sheets into a sealed box at the end of each of their sessions and notified the researchers by a note in the box when they had completed all sets. In one final session, each judge was given the entire set of 80 books, arranged according to her ranking, and was asked to make any final adjustments in rank order that became apparent when viewing books across all sets. No tied ranks were allowed, and all judgments were made independently. The entire procedure was run without any contact between the experimenters and the judges. Judges were paid by the hour, plus a bonus for prompt and responsible completion of the task.

VI. TEXTUAL VARIABLES SELECTED AS PREDICTORS OF READABILITY

In this section we present the rationale underlying our selection of textual variables for predicting readability, and the process of obtaining values for the variables.

Practical considerations affecting selection of variables

The orientation of our study was to seek a practical formula (i.e. one that could be applied easily) which would predict the relative difficulty of children's trade books. The computer provides an efficient means of sorting and counting textual elements and combining them in various ways; whereas hand coding syllables, clauses, word types and/or sentence constructions is very time consuming and not entirely reliable. We therefore restricted our search for predictive textual variables to combinations of measures which could be obtained from direct input of the text as it is written in the books without prior hand coding of syllables, clauses, etc. This specification may have resulted in bypassing some strong variables, but we considered efficienty of prime importance.

Data Text[1] (Armor & Couch, 1972) is a data processing routine which accepts raw text as data. It can count the frequencies of all words and punctuation marks in the text and make separate counts for any words that can be listed and entered as "concepts." Consequently, form classes of words of a finite and predictable size (e.g., pronouns, conjunctions, etc.), terminal punctuation marks (e.g. periods, exclamation points, and question marks), and specific vocabulary lists are possible concepts which can be measured and combined as variables.

TABLE 7

Means, Standard Deviations, and Range of Five Judges' Summed Ranks for Each Set of 20 Books

Set	$\bar{X}$	SD	Range
I (20 books)	204.6	115.9	30 - 391
II " "	206.4	124.5	8 - 385
III " "	194.2	117.4	7 - 390
IV " "	204.9	112.2	15 - 386
Four sets combined, Total of 80 books	202.5	115.4	7 - 391

Correlations of the judges' summed ranks with other data on these same books in their original sets of 20 are presented in Table 8. The generallyhigh correlations between the ranked difficulty of the books and independent cloze and oral reading errors supports use of the ranking data in developing a readability measure that will not be substantially different had cloze data been used instead. At the same time, the very high interrater reliability (.98) has produced data that are more stable than the original cloze scores.

- - - - - - - - - - - - - - -
Insert Table 8 about here
- - - - - - - - - - - - - - -

TABLE 8

Pearsonian Correlations between Judges' Summed Ranks of the 80 Books and Independent Difficulty Data Based upon Cloze Errors and Oral Reading Errors.

	Set I	Set II	Set III	Set IV	Average Correlation
Cloze Errors	.884	.810	.923	#	.872
Oral Reading Errors	.928	.799	.808	#	.845

These data not collected for Set 4.

N = 20

$p < .01$, $r = .492$

FOOTNOTES

[1]Some of the books in the different sets were, in part, difficult for different reasons. For example, the abstract concepts in Where the Wild Things Are (M. Sendak) in Set III compared with the somewhat "foreign" content of Anatole and the Robot (E. Titus) in Set I.

[2]The Spache formula has now been revised (Spache, 1974) and may be used for the primary grades.

VI. TEXTUAL VARIABLES SELECTED AS PREDICTORS OF READABILITY

In this section we present the rationale underlying our selection of textual variables for predicting readability, and the process of obtaining values for the variables.

Practical considerations affecting selection of variables

The orientation of our study was to seek a practical formula (i.e. one that could be applied easily) which would predict the relative difficulty of children's trade books. The computer provides an efficient means of sorting and counting textual elements and combining them in various ways; whereas hand coding syllables, clauses, word types and/or sentence constructions is very time consuming and not entirely reliable. We therefore restricted our search for predictive textual variables to combinations of measures which could be obtained from direct input of the text as it is written in the books without prior hand coding of syllables, clauses, etc. This specification may have resulted in bypassing some strong variables, but we considered efficienty of prime importance.

Data Text[1] (Armor & Couch, 1972) is a data processing routine which accepts raw text as data. It can count the frequencies of all words and punctuation marks in the text and make separate counts for any words that can be listed and entered as "concepts." Consequently, form classes of words of a finite and predictable size (e.g., pronouns, conjunctions, etc.), terminal punctuation marks (e.g. periods, exclamation points, and question marks), and specific vocabulary lists are possible concepts which can be measured and combined as variables.

Guidance from Past Research in the Selection of Variables

Over the years, readability researchers have commonly predicted the relative difficulty of reading passages by combining measures of vocabulary or word difficulty found in the passages with measures related to syntactic complexity or sentence structure. Researchers have studied other factors: measures of conceptual difficulty and abstraction level, the number of personal references in the selection, redundancy, organization of the text, idea density, and human interest. Methods used to measure factors other than vocabulary and sentence structure have not proven to be as reliable and work on them has been discouraging. As indicated by Chall (1958, p. 54), these additional factors usually can be shown to be related to the vocabulary and sentence factors and thus do not add much power to the prediction formulae (i.e. the size of the multiple correlation coefficient) once vocabulary and sentence factors are used.

Measures of vocabulary. The earliest work in readability (Lively and Pressey, 1923) relied on the word frequency counts in Thorndike's Teacher's Word Book (1921), and historically vocabulary has been an important variable in many studies. Vogel and Washburn (1928) used the Thorndike list and Dale and Tyler (1934) created a list of 769 words found in both Thorndike's first 1000 and the word list of the International Kindergarten Union (1928). Dolch devised what he called a "combined word study list" (1928), Lorge used the list of 769 used earlier by Dale and Tyler (1939 and 1948), Dale and Chall (1948) used another list constructed by Dale of 3000 words known to be familiar to fourth grade students, and Spache (1953) as well as others used the Dale 769 list, and later the Stone which was a revised Dale 769 (Stone, 1957). Tribe (1956) appears to be the only one who used a list

(Rinsland's Basic Vocabulary of Elementary School Children, 1945). resulting from children's own work.

A second word measure often used is the number of different words or some other measure of vocabulary diversity (Washburne and Morphett, 1938; Gray and Leary, 1935; Dolch, 1928).

Various measures related to word length have also been used in other formulas. Average number of letters per word (McClusky, 1934), the number or percent of polysyllabic words (Johnson, 1930; Wheeler and Smith, 1954; Gunning, 1952; etc.), the percent of monosyllabic words (Farr, Jenkins, and Patterson, 1951), and the number of syllables per 100 words (Flesch, 1950; Fry, 1967) have all been studied. Finally, letters per independent clause, and letters per sentence appear in Bormuth's 1964 study along with syllables per word, per independent clause and per sentence.

A somewhat different approach to vocabulary measure was taken by Lewerenz (1930) who differentiated between simple Anglo-Saxon words and difficult technical and special meaning words of Greek and Roman derivation. He also studied the assignment of words to categories of "easy" and "difficult" according to their initial letters Lewerenz, 1929).

Syntactic complexity. Readability formulas traditionally reflect the difficulty of sentence structure. Although few researchers have used counts of specific sentence construction as measures of difficulty, most have tallied factors such as sentence length and/or specific parts of speech which are related to sentence complexity.

Words per sentence has constantly recurred in formulae from the early studies of Gray and Leary (1935) and Lorge (1939) through the period of the popular readability formulae of Flesch (1943),

Dale and Chall (1948), Spache (1953) and others, to Tribe's study of 1956 and Fry (1969). Ojemann (1934) and Dale and Tyler (1934) counted simple, complex, and compound sentences in their studies and Vogel and Washburne (1928) as well as Washburne and Morphett (1938) used the number of simple sentences in 75 sample sentences of a 1000 word passage into their readability formulae.

Counts of specific parts of speech thought to reflect sentence complexity have varied considerably, with only prepositions as a frequently recurring measure. Vogel and Washburne (1928) and Ojemann (1934) counted prepositions in their sample passages. Dale and Tyler (1934) used clauses and prepositional phrases, while Gray and Leary (1935) and Lorge (1939) counted prepositional phrases.

Bormuth (1964) also made counts of various parts of speech in a way that is a significant departure from traditional readability work. He hand-coded all words according to form class and/or part-of-speech and then took advantage of computer technology to research language units in much greater depth than had been done previously. He made a systematic search for linguistic variables which yielded high correlations with his measure of passage difficulty.[2] Among the variables included were many unusual ratios of the eight parts of speech of traditional grammar and five of the form classes of modern linguistics.

The multiple correlations reported in the Bormuth study are most impressive. The difficulty of coding words as to their part-of-speech or form class and the difficulty of coding independent clauses severely limits the possibility of including these very strong variables into a readability formula that might easily be applied to thousands of children's books. However, the relative strength of his variables provided guidance for our own selection of variables.

The Textual Variables Selected

After researching the literature cited above and seeking ways to simplify and/or strengthen relevant variables by calculating second order variables, we decided upon a set of variables which would be combinations of measures easily made by computer and which might be expected to be predictive. Data Text (Armor and Couch, 1972) provided the means for defining the "concepts" and the variables were ratios of the concept counts which were calculated by a separate computer routine. The variables fall into five categories: (1) variables related to vocabulary, (2) variables related to syntactic structures, (3) variables reflecting the extent of personal references, (4) variables which combine syntactic complexity and personal references, and (5) variables of format determined by the publishers which we have called "summative variables." Each category is described below.

(1) Vocabulary. Word length has traditionally been found to be a strong variable: Bormuth's data (1964) suggested that a useful division might be made by placing words of five letters and shorter into the category of "easy" words and six letters and longer into the "difficult" category. We obtained data on the number and frequency of one-letter words, two-letter words, three-letter words, etc., from the 40 books in Sets I and II. Then separate ratios of all words over 5 letters per total number of running words, all words over 6 letters per total number of running words, etc., were computed for each of the 40 books. The 20 books in each set were assigned two rank orders: one according to the cloze score and one according to the percent of words over five letters in length. Rank order correlations between the two were computed. The same was done using words over six letters to rank

the books and another correlation computed, etc. These correlations are given in Table 9. The categories beyond nine letters were not computed as the frequency of such words was negligible in most books.

Insert Table 9 about here

There is little difference between the correlations for the different word lengths. Based partially on intuition, on knowledge of the actual number of words found in the selections in these four categories, and on the relative stability of the correlations between sets, a decision was made to use the measure of words over seven letters in the variables designed to give word difficulty measures. We computed the following two variables: number of words over seven letters per total number of running words and number of words over seven letters per number of sentences (variables 1 and 2, Table 10).

Familiarity is understood to be a strong psychological variable and a passage densely infused with familiar words should be easily understood, all other things being equal. Past research has consistently yielded significant correlations between the relative frequency of familiar words in a passage and a reader's understanding of that passage. The 769 words used by Dale and Tyler (1934) (i.e. those common to the first 1000 on Thorndike's list and also appearing in the International Kindergarten Union list) seemed most appropriate to our study of younger children's reading. That list was expanded (i.e. plurals, past tense of verbs, possessives, etc., were added) according to the rules given in Dale and Chall (1948) for expanding the list of 3000.[3] The result was a list of about 2000 words (the 769 plus their derivatives) to be entered as a

TABLE 9

Rank Order Correlations between Cloze Score Ranks for the Books and the Percent of Words Containing More than the Specified Number of Letters

	Cloze with % of Words over 5 Letters	Cloze with % of Words over 6 Letters	Cloze with % of Words over 7 Letters	Cloze with % of Words over 8 Letters
Set I (N=20)	.60	.71	.79	.77
Set II (N=20)	.77	.81	.72	.65
Sets I and II Combined (N=40)	.72	.73	.76	.73

"concept" for computer analysis. These measures were used to compute two additional word variables: number of Dale words per total number of words and number of Dale words per total number of sentences for each book (variables 3 and 4, Table 10).

We also considered devising a new and much shorter list of highly frequent words which might have a more practical application than the long Dale list and yet be predictive. Frequency ratios for the most frequent words in the 40 books in Sets I and II were computed. Correlations with cloze scores on these same books indicated that (1) there were individual frequent words with high positive correlations and some with high negative correlations with cloze ratings[4] (i.e. some frequent words occurred most frequently in the difficult books and some occurred most frequently in the easier books) and (2) the correlations were noticeably unreliable across the two sets of books for many of the words. Refining of the list according to some theoretical rationale and searching for some consistencies and/or logic to the differences in the two sets of words could be useful and productive, but we did not pursue that work for the current study.[5]

(2) Syntactic complexity. Sentence length is a variable which has been used often to reflect difficulty of sentence structure. The rationale offered long ago by Dale and Tyler (1934) remains appropriate: "It seems likely, when sentences are used which involve suspension of one's judgment as to the outcome until the entire sentence has been covered, that the difficulty would be increased." (p. 397) In order to include the number of words per sentence (variable 5, Table 10) we obtained counts of the number of words in each book and the number of sentences in the book.

whereas they represented only 7 percent of different words used by adults. She also reported that when the percentage of pronouns relative to the total number of words used was computed, it was higher for the children (21.5%) than for adults (18.8%).

If a division of pronouns could be made that would place pronouns used more often in speech in one category and others in a second category, such a categorization could, theoretically, relate to the comprehension of written materials. There are several possible ways to divide pronouns in order to clarify if such a distinction really exists. For instance, third person personal pronouns might be less "personal" and it might be more useful to categorize them with the impersonal pronouns. Rodgers (1967, p. 6) states, "Frequency studies have shown... that first and second person pronouns dominate spoken English whereas third person pronouns dominate written English." Our own observations of the books in our collection substantiated the notion that dialogue and simple, direct writing often contains more first and second than third person personal pronouns.

Preliminary separate counts were therefore obtained on first and second person personal pronouns (or "personal personal pronouns" as we dubbed them), third-person personal pronouns, all personal pronouns, and impersonal pronouns to be used in calculating variables 9 through 14 (Table 10). These variables are ratios obtained by dividing the counts of pronouns by the number of words in the book and by the number of sentences in the book. Perhaps some ratio of the categories of pronouns themselves relative to each other would provide a better key to the personal-talking style of writing. In order to investigate such a possibility, variables 15, 16 and 17 (Table 10) were included. It was not expected that all of these variables would be useful, but assessing

their relative strengths was important if, indeed, our intuitions regarding personal references had any validity at all.

(4) Combined syntactic complexity and personal references. One variable described under category (2), personal pronouns/conjunction, really combines both syntactic complexity and personal reference. If we were to add personal personal pronouns/conjunction also, comparative data on the two variables would give further evidence for the validity of separating out personal personal pronouns. Variable 19 (Table 10) was therefore included.

(5) Summative variables. Books written for beginning readers are designed so that they appear extremely easy. The number and kind of pictures that are included and the size of print are immediately obvious cues to even the very casual observer. Such variables are not easily measured, but some format considerations do lend themselves to reliable and practical measurement. The number of words per page is a second order variable (our variable 20, Table 10) that reflects both print size and the ratio of pictures to print. In and of itself, it is a variable one assumes is consciously manipulated by publishers to make a book appear more or less difficult. It may also be the case that a given story printed with fewer words per page is easier to read for some underlying psychological reason having to do with the expectations of the reader.

Often books for the less experienced reader are, by design, shorter than those for more advanced readers and so the actual length of the book was considered in variables 21 and 22 (Table 10).

Another formal consideration was prompted by the fact that basal readers used in teaching beginning reading usually do not have sentences

begun on one page and carried over to the next page. Do publishers of children's trade books control the same variable? If so, is there a relationship between the number of pages that have sentences carried over and the difficulty of the material? We decided to investigate this variable also (variable 23, Table 10).

Words per page, number of words in the book, number of pages in the book, and the number of sentences carried over to another page per total number of pages are all controlled by the publisher. Each was included based on the assumption that publishers do manipulate these variables to reflect their own judgments as to the difficulty of the stories. In that sense they are "summative" variables.

Insert Table 10 about here

Computer Analysis of the Text in Children's Books

The textual variables described above were computed for all 80 books on the entire book and also separately for the shorter passages read by the children in 60 of the books (Sets I, II and III of 20 books each). The text was directly keypunched[6] as printed in the books.

These data were then batch processed on the Data Text Program (Armor & Couch, 1972). Output from this program was an alphabetical listing and word count for each of the (books) submitted and separate tabulations for each book of the "concepts" discussed as potential measures in the previous section: impersonal pronouns, personal personal pronouns, third-person personal pronouns, prepositions, conjunctions, and the 2000 words on the expanded 769 Dale list as well as the number of words, number of pages, number of sentences, number of commas, gaps,

TABLE 10

Final List of Variables

1. Words over 7 letters/word
2. Words over 7 letters/sentence
3. Dale 769 words/word
4. Dale 769 words/sentence
5. Words/sentence
6. Prepositions/sentence
7. Conjunctions/sentence
8. Commas, colons + semicolons/sentence
9. Personal personal pronouns/word
10. Personal personal pronouns/sentence
11. Personal pronouns/word
12. Personal pronouns/sentence
13. Impersonal + third-person personal pronouns/word
14. Impersonal + third-personal personal pronouns/sentence
15. Impersonal pronouns/personal pronouns
16. Impersonal pronouns/personal personal pronouns
17. Impersonal + third-person pronouns/personal personal pronouns
18. Personal pronouns/conjunction
19. Personal personal pronouns/conjunction
20. Words/page
21. Number words in book
22. Number pages in book
23. Carried-over sentences/page

colons and semicolons, and the number of sentences carried over from one page to another in the book (see Appendix I).

All of the variables were ratios of one of these measures to another (see Table 10) and they were calculated by a separate computer routine.

In addition to the value of each textual variable on each book, we computed the mean for the textual variables on the Formative and Validative Sets of books (i.e. a mean for the 50 Formative books and a mean for the 30 Validative books); these are given in Table 11. The similarity of the values of these variables on the two sets reinforces the notion that the two sets are, indeed, representative of the same universe of books.

- - - - - - - - - - - - - - - - - -

Insert Table 11 about here

- - - - - - - - - - - - - - - - - -

TABLE 11

Means and Standard Deviations for Textual Variable and Criterion Variable for Two Subsets of the 80 Books. 50 Formative Books and 30 Validative Books.

Variables	50 Formative Books Mean	50 Formative Books S. D.	30 Validative Books Mean	30 Validative Books S. D.
1. Words 7^{+} letters/words	.0570	.0411	.0413	.0303
2. Words 7^{+} letters/sentences	.6163	.5303	.4412	.3829
3. Dale/word	.7830	.0684	.8063	.0639
4. Dale/sentence	7.8182	3.5131	7.5220	3.1282
5. Words/sentence	9.9848	4.7092	9.3637	3.8963
6. Prep./sentence	1.1202	.6177	1.0626	.5928
7. Conj./sentence	1.6272	2.5739	1.7288	2.2541
8. , : ; /sentence	.6380	.4130	.4649	.2214
9. Personal personal prn./word	.0465	.0365	.0591	.0497
10. Personal personal prn./sentence	.4432	.5801	.4645	.3509
11. Personal prn./word	.1050	.0344	.1161	.0450
12. Personal prn./sentence	1.0075	.5462	1.0358	.4759
13. Imp. + 3rd person personal prn./word	.0942	.0220	.0974	.0300
14. Imp. + 3rd person personal prn./sentence	.9237	.3882	.9364	.5519
15. Imp. prn./personal prn.	.3808	.2375	.4733	.5219
16. Imp. prn./personal personal prn.	2.2522	3.7268	4.8450	15.6515
17. Imp. + 3rd person personal prn./personal personal prn.	6.4544	11.8927	4.1840	8.8285
18. Personal prn./conj.	2.7598	2.0343	3.1480	2.9050
19. Personal personal prn./conj.	1.2032	1.1652	1.7060	1.8844
20. Words/page	48.4380	48.5387	47.0500	51.8637
21. Words in book	1558.8400	1614.2470	1330.9670	1193.6740
22. Pages in book	34.0600	13.4959	33.2000	12.6202
23. Sentences carried over/page	.1198	.1878	.1600	.2038
24. Judges' summed ranks	209.1400	115.3660	191.4330	116.6470

FOOTNOTES

1
Data Text also provides the statistical routines for correlation and multiple regression analysis which were necessary for the study.

2
"Passage difficulty" in the Bormuth study is the average word difficulty for each passage, computed directly from the proportion of subjects who guessed each word, or cloze item, in the passage correctly.

3
See Appendix F.

4
See Appendix G for the words and their correlations with cloze scores.

5
It is interesting to note in passing that words with high positive correlations as well as words with high negative correlations with the cloze scores appear on the Dale 769 list. Also some of the Dale 769 words are positively correlated in one set and negatively correlated in another set. This phenomenon could be a function of our set of books which are, in general, written for younger children than the books used in the Dale and Tyler study. In any event, our data do cast doubt on the reliability of word lists based on frequency only without regard to other attributes of the individual words (e.g. their form class or function).

6
Appendix H shows the coding of punctuation marks and other rules for keypunching.

VII. STATISTICAL DEVELOPMENT OF THE READABILITY FORMULAE

Given the textual variables selected for potential indices of readability, and the criterion variable of reading difficulty provided by judgmental ranking of the 80 books as well as cloze error data on 60 of the books, readabilty formulae were generated using multiple regression techniques. Three types of analyses are reported:

1) Correlations between the textual variables and the criterion of judges' ranking of book difficulty.

2) Regression equations based upon the textual variables predicting ranked difficulty of the books.

3) Validation analyses from data based upon the predicted difficulty of 30 books that were not a part of the Formative Set used for developing the readability formulae. These data are extended to the original performance-based criterion of cloze and oral error scores.

Correlational Analyses

Pearson product moment correlations were performed between the 23 textual variables and the criterion of ranked difficulty of the books. As shown in Table 12, these correlations were performed separately for the 50 books selected for the Formative Set and the 30 books selected for the Validative Set. For most of the variables with significant correlations in the Formative Set there are also significant correlations with the Validative Set. This is true especially for the variables with higher correlations, which are those most likely to appear in a multiple regression equation.

- - - - - - - - - - - - - - - - - - -
Insert Table 12 about here
- - - - - - - - - - - - - - - - - - -

TABLE 12

Coefficients of Correlation Between Textual Variables and Judges' Summed Ranks[1]

Textual Variable	Correlation between variables & judges' summed ranks on 50 books in Formative Set.	Correlation between variables & judges' summed ranks on 30 books in Validative Set.
1. Words 7+ letters/word	.516**	.787**
2. Words 7+ letters/sentence	.645**	.819**
3. Dale 769 words/word	-.545**	-.503**
4. Dale 769 words/sentence	.422**	.513**
5. Words/sentence	.499**	.603**
6. Prepositions/sentence	.523**	.579**
7. Conjunction/sentence	.272	.319
8. , : ;/sentence	.431**	.765**
9. Pers. pers. prn./word	-.392**	-.437*
10. Pers. pers. prns./sentence	-.077	-.123
11. Pers. prns./word	-.393**	-.321
12. Pers. prns./sentence	.192	.327
13. Imp. + 3rd pers. prns./word	.147	.253
14. Imp. + 3rd pers. pers. prns./sentence	.632**	.545**
15. Imp. prns./pers. prns.	.100	-.112
16. Imp. prns./pers. pers. prns.	.219	-.032
17. Imp. + 3rd pers. prns./pers. pers. prns.	.195	.306
18. Pers. prns./conjunction	-.393**	-.095
19. Pers. pers. prns./conjunction	-.526**	-.307
20. Words/page	.685**	.745**
21. Number words in book	.585**	.698**
22. Number pages in book	-.295*	-.500**
23. Carried-over sentences/page	.260	.458*

	*	**
P <	.05	.01
r =	.282	.365

An examination of Table 12 shows two major types of variables, those that might be termed "linguistic" (e.g. 1-19) and those which we call "publisher" variables, that are determined by the book-making preferences of the publisher (20-23). These "publisher" variables, such as words/page, are clearly not unidimensional linguistic measures, but represent the summative effects of many textual variables. For example, words/page probably depends upon: word length (variables 1 and 2), Dale words (variable 3), prepositions/sentence (variable 6), and the publisher preferences about illustrations, type size and words per page. So, although words/page correlates highly with the criterion variable, and may be of potential practical value in predicting book difficulty, it is not very interesting linguistically and may not be useful for those books where publishers do not adhere to current practices of book making.

Regression Analyses

Using the above textual and criterion variables the step-wise multiple regression procedure of Data Text was used to develop a variety of regression equations predicting difficulty of the books.

We here report two readability formulae: Type "P" (for Publisher) based upon a regression analysis that included the publisher variables (20-23); and Type "L" (for Linguistic) which excluded the publisher variables.

Results. The first step-wise multiple regression run, all variables submitted, is summarized in Table 13. This resulting regression equation is hereafter referred to as "Formula P" because it includes publisher determined variables.

- - - - - - - - - - - - - - - - - -

Insert Table 13 about here

- - - - - - - - - - - - - - - - - -

TABLE 13

Stepwise Regression Analysis of the Correlations Between Book Difficulty (Judges' Summed Ranks) and Textual Variables including all Variables for 50 Books in Formative Set.

Step	Variable Entered	R	R^2	F	P
1	(#20) Words/page	.6850	.4692	42.43	.000
2	(#14) Imp. + 3rd person pronouns/sent.	.8185	.6699	47.68	.000
3	(# 3) Dale/word	.8582	.7364	42.84	.000

Words per page, a publisher determined variable, rather than any of the linguistic variables, had the highest correlation with the criterion variable and was therefore the variable selected for the first step. Impersonal pronouns plus third person personal pronouns per sentence was the variable selected by the computer at Step 2 with a noticeable increase in the multiple correlation. Step 3 increased the correlation only slightly. The difficulty of computing the variables for such a small increase led to the decision to use only the first two variables in later computing the predicted difficulty scores for the books.

The regression equation or Formula P is: 5.447 (words/page)+ 469.4 (impersonal pronouns + 3rd person personal pronouns/sent) + 17.567 with a multiple correlation of .819 with reading difficulty as determined by judges' summed ranks on 50 books.

The results of a second regression analysis using only the linguistic variables, eliminating the summative variables (words per page, words in book, pages in book and sentences carried over to a second page per page) are given in Table 14. This regression equation is referred to as "Formula L" because it was derived on linguistic variables only, excluding all summative variables.

- - - - - - - - - - - - - - -
Insert Table 14 about here
- - - - - - - - - - - - - - -

In this analysis, a variable which probably reflects both vocabulary difficulty and sentence complexity, words over 7 letters long per sentence, had the highest correlation with the criterion variable and was therefore selected for the first step. Again the second step involved variable #14, impersonal plus 3rd person personal pronouns per sentence. The third step increased the correlation with criterion only slightly and was therefore not used in computing predicted difficulty scores for the books.

TABLE 14

Stepwise Regression Analysis of the Correlation between Book Difficulty (Judges' Summed Ranks) and Textual Variables for 50 Books in Formative Set, Excluding Summative Variables.

Step	Variable Entered	R	R^2	F	P
1	(# 2) words 7^+ letters/sent.	.6449	.4159	34.18	.000
2	(#14) Imp. + 3rd person pronouns/ sent.	.7181	.5156	25.02	.000
3	(# 3) Dale/word	.7420	.5506	18.78	.000

The regression equation or Formula L is: 418.8 (words over 7 letters/sent) + 388.4 (impersonal pronouns + 3rd person personal pronouns/sent) + 46.352 with a multiple correlation of .718 with reading difficulty as determined by judges' summed ranks on 50 books.

Validation Analyses

The validity of Formulas P and L can be assessed by looking at the correlation between predicted difficulty of the books and their difficulty as shown by the following criteria:

1) Judges' summed ranks of the 30 books in the Validative Set.

2) Cloze scores of the three separate sets of 10 validative books each.

3) Oral error scores of the three separate sets of 10 validative books each.

Results are shown in Table 15. There is reasonably good consistency in the correlations across the different criteria and sets of books, considering the relatively small n's involved in the Validative Sets (n = 10).

- - - - - - - - - - - - - - - - - -
Insert Table 15 about here
- - - - - - - - - - - - - - - - - -

Applying the Readability Formulae

The two regression equations used in computing reading difficulty scores for the books are:

Formula P

$$X_P = 5.447X_2 + 469.4X_3 + 17.567$$

Formula L

$$X_L = 418.8X_4 + 388.4X_3 + 46.352$$

X_p = reading difficulty score including summative (or publisher determined) variables

X_L = reading difficulty score including linguistic variables only

X_2 = the number of words in the book divided by the number of pages of text in the book

X_3 = the number of impersonal pronouns and third person personal pronouns in the book divided by the number of sentences in the book

X_4 = the number of words more than seven letters long contained in the book divided by the number of sentences in the book (include every instance of the word, not merely the number of different words over seven letters long)

Two predicted reading difficulty scores, one computed on each of the two regression equations, are given in Appendix A for all 80 books; 50 formative and 30 validative.[2] The range of the reading difficulty scores using Formula P is 158 for the easiest book[3] to 1881 for the most difficult[4] (error of estimate = 67.68); while the range of reading difficulty scores using Formula L is 142 for the easiest book[5] to 1663 for the most difficult[6] (error of estimate = 81.98). These predicted scores were then plotted against the judges' summed ranks and that graph is presented in Appendix K.

TABLE 15

Pearson Product-Moment Correlations of Scores Given to Books by the Two Readability Formulae with Judges' Summed Ranks, Cloze Scores, and Oral Error Scores

Readability Formula	Books	Correlation with: Judges' summed ranks	Children's cloze scores	Children's oral error scores
Formula P				
	Formative Set (n = 50)	.816***		
	Formative Books: Set I		.714*	.767*
	" " Set II		.562	.516
	" " Set III		.813**	.732*
	Validative Set (n = 30)	.815***		
	Validative Books: Set I (n = 10)		753*	.812**
	Validative Books: Set II (n = 10)		.543	.720*
	Validative Books: Set III (n = 10)		.629	.450
Formula L				
	Formative Set (n = 50)	.716***		
	Formative Books: Set I		.628	.717*
	" " Set II		.597	.529
	" " Set III		.842**	.870**
	Validative Set (n = 30)	.713***		
	Validative Books: Set I (n = 10)		.835**	.906**
	Validative Books: Set II (n = 10)		.641*	.817**
	Validative Books: Set III (n = 10)		.540	.428

	n=10	n=20	n=30	n=50
*p < .05	.497	.360	.296	.231
**p < .01	.658	.492	.409	.322
***p < .005	.708	.537	.449	.354

(e.g., for cases where n=10, a correlation of .497 has a p < .05, a correlation of .658 has a p < .01, etc.)

FOOTNOTES

1
Intercorrelation matrices of all variables are given in Appendix J.

2
The judges' summed ranks are also given in Appendix A.

3
Where is Everybody by R. Charlip

4
The Adventures of Beetlekin by Dulieu

5
Ten Apples Up on Top by T. LeSieg

6
Casey at the Bat by E. Thayer

VIII. DISCUSSION

In this section we would like to discuss the results, which are encouraging in several respects, and to make a case for training teachers and librarians in the art of rank ordering books by inspection. We also present limits to the interpretation of the study and a discussion of the relationship of our formulae to other formulae.

Results

Originally we undertook this study because no existing readability formula had been developed or validated on young children's literature or trade books. The formulae developed in this study yield significant correlations with the criterion of judges' summed ranks, and validity of the formulae on material that had not been used in devising them has been demonstrated. The correlations indicating validity to outside criteria, children's cloze scores and children's oral error scores are within the range expected for such measures. (See Table 15, page 37) Given the practical limitation of studying only those variables which could be easily subjected to computer analysis without hand-coding, the variables in the two formulae account for a considerable share of the variance: approximately 67% and 52% respectively.

Our recommendations for the use of the two formulae are quite straightforward. Formula P might be used by educators and librarians to put books which have maintained the format determined by the publisher into relative positions of difficulty. This formula is limited in its generalizability by the fact that it is dependent upon the publishing practices and craft that go into the production of children's trade

books. As children get older, material written for them is less subject to differentiation on the variable of words per page and for that reason perhaps one should not consider the formula valid beyond about the third or fourth grade level. An additional restriction is, of course, that it must not be used for literature that is no longer in its original format. Stories taken from a book and reinserted in another context lose their original count of words per page and unless one can get back to that original count, the validity of Formula P is doubtful.

Formula L is more generalizable, though somewhat less powerful. It should prove useful to researchers and educators in estimating the readability of a variety of materials drawn from the field of children's literature for readers in the primary grades. Planned systematic instruction in reading using the resource of children's literature, should be possible when this formula is used to rank order the books. More specific instructions for use of this formula will be published, including a nomograph for easy application of the formula. For those with access to a computer, a count of impersonal and third person personal pronouns, the number of sentences in the book and the number of words greater than seven letters in length is easily accomplished, as is the actual mathematics for the formula.

The skill of rank ordering children's trade books on the basis of inspection has been shown by this study to be easily acquired and quite reliable. The procedures we used are described on page 52 and our data indicated very high correlations between the judges' ranking and the children's cloze and oral error scores (see Table 8, page 61). Potentially, such correlations could be much higher than those we obtained if the cloze measure could be improved upon.

In terms of reliability of the judges' ranking, one could not ask for more than the obtained results. It remains to be seen whether such results can be replicated by other judges, with other books. The obtained correlations with the independent criteria suggest validity at least as great as for conventional readability formulas. Thus techniques of judgment appear to hold promise for further applications to research and practice (see Klare, 1974, p. 64 also). An indication that the judges' ranks correlate more strongly with the criterion variables than extant readability formulas is found in a comparison of these rankings with our own readability formulae. The correlations of our two formulae with the cloze scores on the formative set (.816 and .716) and the correlations of the two formulae with the cloze scores on the validative set (.815 and .713) were not quite as strong as the correlations of the judges' ranks with cloze scores on Set I (.884), Set II (.810), or Set III (.923) (see Table 8, page 61 and/ Table 15, page 87). It remains an empirical question as to how well the revised Spache (1974) formula, for instance, would predict, but it may well be that there are untapped variables taken into account by a sensitive human observer, which are as yet not incorporated into readability measures. Or perhaps the human observer provides a more sensitive weighting of existing variables than can be obtained from presently used regression techniques.

The sensitivity of the human observer is reinforced by the inspection of certain data where the formulae yielded quite low correlations. We might, for instance, consider "abstractness" a variable more subject to human observation than computer analysis. Sam, Bang and Moonshine by E. Ness is perceived by judges to be more difficult in relation to some other books in Set III than Formula L predicts and, in fact, it is

more difficult according to the cloze performance. This abstract difficulty, we can hypothesize, was also perceived by the publishers of the book, for Formula P predicts the difficulty more accurately than Formula L. We could infer that their chosen format reflects what they perceived to be greater difficulty. Similarly, but for quite different reasons, humans perceive greater difficulty for Fox in Sox than the formulae. The alliteration and rhyming in this book by Dr. Seuss is quite sophisticated for a young reader reading the book for himself. Qualities to which human beings are more sensitive, will likely always fall outside the range of a practical formula. Such issues should be approached in future research.

For the present, a practical approach to ranking children's literature according to its expected readability for young children is to be found in training educators and librarians to assess relative difficulty in a manner similar to that described in this report.

It should be noted that the results of the judges' ranking were obtained with brief training of the judges and about 12 hours of effort by each judge. This amounts to a total time investment of about 45 minutes per book for the five judges, or nine minutes per judge per book, well within reasonable limits for practical applications of the judgmental technique. Perhaps educational practitioners as well as researchers will gain increased confidence in their intuitive judgments, tempered by systematic procedures, from the results presented here.

There is one further practical merit to the present findings. The list of 80 ranked books can provide a reading difficulty scale for the judging of other childrens' trade books, and the set of books themselves provide the core collection of a difficulty-scaled corpus of reading

matter for use in teaching at the primary level and in research.

Some Limitations

The discussion presented below of our criterion measures will remind the reader to restrict his expectations of the formulae. This discussion is then followed by a rationale for our decisions not to assign grade-levels to the books.

Criterion measures. In an earlier section of this paper we discussed the range of children's behaviors that might be appropriate as a criterion measure for research on readability of young children's trade books and advanced an argument for the use of the cloze procedure as a criterion variable in our research.

Children's cloze scores were collected for three sets of 20 books each with the intention of using them as a criterion variable. However, preliminary correlational analyses of textual variables with the cloze scores on the three sets indicated that many strong variables were unstable; i.e., a variable which had a high correlation with the cloze scores on one set of 20 books might have much lower correlation on one of the other sets of 20 books. The advantages of using a larger set of books in developing the formula were obvious but the technical difficulty of obtaining cloze scores that would be comparable across all 80 books was insurmountable. Therefore we used jusges' summed ranks which related each book to the entire set of 80, and the set of 80 was divided into a formative set of 50 for developing the formula and another set of 30 for validating it.

Although the correlation of the judges' summed ranks and the cloze data are very high (see Table 8, page 61), we would like to remind our readers that the formulae were developed on adults' perceived difficulty

of materials and not on children's behavior as originally planned. However, certain advantages accrue from that fact. The judges' summed rank for each book is based on the entire book, thus making it more representative of that book than the cloze scores based on only a 100-word passage. The construction of cloze passages in children's books needs considerable refinement. Passages to be read as samples from the books need to be selected in a more sophisticated manner so that one would know whether or not each accurately reflects the general level of difficulty for a given book (Clymer, 1959).

The discussion of our problems in generating a valid criterion measure for our research and our experience and observations in collecting data from judges and from children prompt us to note a restriction on what to expect from either formula. Many books vary considerably in difficulty from one part to another. Although one may assess the average difficulty of the book by applying either formula to the entire contents of the book, one may not assume that a child will be able to read all parts of it equally well. There may be parts of that book which are, in effect, much more difficult and unexpected problems may arise for the young reader. The same caution should be made for difficulty levels assigned by other formulae as well. Add to this the complicating factor of "interest," peculiar to each individual child, and we are made keenly aware of how limited our schemes for assessing difficulty are. All schemes are approximations and should be treated as such. No criterion measure utilized to date, to our knowledge, is capable of precisely assessing the degree of difficulty a given individual child will have in understanding a given book.

"Grade-Level." Grade level assignments have purposely not been

computed, nor have tables for transforming readability scores into grade levels been constructed. Although some may construe this fact as a "limitation," we view it as such only in the milieu of the current pre-occupation with standardized test norms. Grade levels merely reflect a mean or median level of performance on a standardized test given to a group of children at that level. Therefore, assigning a grade level to a book does not indicate which children in any particular grade or classroom should be reading that book even if we know his score on a standardized test. (See Auerbach, 1971, for data relating "readability" to standardized achievement tests.) Further, more and more ungraded classrooms with multi-aged groups of children are appearing on the educational scene at the primary level and a considered argument can be made for making literature available to the children without grade level designations.

What is needed in the schools is a means by which a child might have access to a very wide range of books at an appropriate level of difficulty, written by different authors in different styles about many topics. He should be able to read and understand the books he selects and then to progress to books assessed as more difficult. A cohesive scheme consisting of a placement test to designate a child's entry point and a designation of the relative difficulty of the books would be necessary to formalize such a plan. Grade levels become irrelevant if not an interference to the enjoyment of literature. However, more work is needed on developing placement instruments and assessing the reading difficulty of a mass of books before an effective system can be built.

Relationship to Other Formulae

Almost certainly, other researchers will be tempted to make

comparisons between the relative difficulty of books as assigned by our formulae and as assigned by other formulae. It would seem particularly appropriate that some of the formulae developed on young children's text books (e.g., Spache, 1974) be compared to ours to see how much difference actually exists when trade books versus textbooks have been used in developing the formula. For practical purposes, educators have assigned levels to simple children's literature for years according to the Spache formula which was devised on material from children's textbooks. The Dale-Chall (1948) formula has been popular for books used beyond the third grade. The Exemplary Center for Reading Instruction in Salt Lake City, Utah, is one group that has scaled books according to these two formulae. It would be interesting to compare levels assigned by the Spache and/or Dale-Chall formula with the 80 rankings established by our formulae and with the cloze scores from this study. The results of such a study might elucidate the practicality of initiating further efforts to rank order a large set of children's literature by means of existing formulae.

Future Research

Several directions for future research are suggested by the present study. As mentioned above, relating our work to earlier readability formulae based on textbooks as opposed to trade books is important. Also, the development of a formal "Placement Test" to accompany the list of 80 books based on our data and observations would make the results more useful to classroom teachers. We would like as well to encourage further investigation with teachers, reading specialists and librarians to help establish a valid and reliable procedure for judgmental ranking of children's literature books.

In addition, data obtained on the correlation of particular variables with the criterion suggest that further research in this area would be fruitful. Two such variables in particular pique our interest. One, variables associated with the "personalness" of the narrative and another, word lists, are discussed briefly below.

The predictive power of the variable on "personalness" resulted in its being included in both formulae. Impersonal pronouns plus 3rd person personal pronouns per sentence results from our categorization of pronouns into "personal-personal" pronouns and other pronouns. Our rationale for categorizing pronouns is presented on page 72 and while we cannot make an argument for "causality" merely because the correlations are high, we believe that more extensive research in this area might prove interesting. It is possible that our division of pronouns has isolated those which reflect a direct style of writing (first and second person personal pronouns or personal-personal pronouns). The remaining category of pronouns (impersonal plus 3rd person personal pronouns) on the other hand, reflects a less direct style of writing. It appears reasonable that the more pronouns of this less personal type there are in a sentence, the more difficult the material should be for the young reader.

There should be considerable pay off in pursuing work on a word list constructed so that each word would meet specific criteria. Study of individual words to determine if proportionately more occurrences of a word correlates positively or negatively with a criterion measure, study of the stability of the correlations across different sets of books, and study of words selected on the basis of some theoretical rationale (other than frequency) are all avenues suggested from the

present research. We were struck by the fact that even considering only a short list of very common words, some of those words correlated positively with cloze scores and some correlated negatively, and even the direction of the correlation was unstable for some words across two sets of books (See Appendix G). It is quite possible that further investigation and development of specialized word lists might result in a formula that would have improved correlations with criteria.

Conclusion

Work in readability remains important, both for the changing trends of instruction for young children and for the ever increasing demand for adult literacy programs. Our work provides a unique contribution in that the formulae were actually developed on children's trade books and some sense of the validity of the formulae on that type of material was demonstrated. We now have the capability of comparing readability levels based on older formulae, which were devised on material from children's textbooks, with reading difficulty scores based on the formulae presented in this study. We also have the capability of assessing the relative difficulty of a host of children's literature books in a relatively efficient manner and of refining a scheme for placing children into this set at an appropriate level of difficulty. Further, we have lent credibility to a procedure of using judges' ranks for assigning difficulty to children's books. We sincerely hope that this work will advance the goal of helping children to increasingly use literature as a source of growth and pleasure.

BIBLIOGRAPHY

Armor, David J. and Couch, Arthur S. Data-Text Primer. New York: Free Press, 1972.

Auerbach, Irma-Theresa. Analysis of standardized reading comprehension tests. An unpublished doctoral thesis, Harvard University, 1971.

Bloomer, R. H. Level of abstraction as a function of modifier load. Journal of Educational Research, March, 1959, 52, 269-72.

Bond, G. L. and Dykstra, R. The cooperative research program in first-grade instruction. Reading Research Quarterly, Summer, 1967, 2 (4), 5-142.

Bormuth, J. R. Relationships between selected language variables and comprehension ability and difficulty. Cooperative Research Project No. 2082, U. S. Office of Education, 1964.

Bormuth, J. R. Comparable cloze and multiple-choice comprehension test scores. Journal of Reading, 1967, 10, 291.

Chall, J. S. Readability: An Appraisal of Research and Application. Columbus: The Bureau of Educational Research, Ohio State University, 1958.

Clymer, Theodore. A study of the sampling reliability of the Spache readability formula in Reading in a Changing Society, International Reading Association Conference Proceedings, Vol. 4, 1959, 245-250.

Dale, E. and Chall, J. S. A formula for predicting readability: instructions. Educational Research Bulletin, February 18, 1948, 27, 37-54.

Dale E. and Tyler, R. W. A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. Library Quarterly, July, 1934, 4, 384-412.

DeLong, V. R. Primary promotion by reading levels. Elementary School Journal, May, 1938, 38, 663-71.

Dolch, E. W. Vocabulary burden. Journal of Educational Research, March, 1928, 17, 170-83.

Dolch, E. W. Graded reading difficulty. (Ch. XXI), Problems in Reading. Champaign: The Garrard Press, 1948, 229-55.

Elley, W. B. Standardized test development in New Zealand: problems, procedures and proposals. New Zealand Journal of Educational Studies, 1967, 2, 63-77.

Farr, J. N., Jenkins, J. J., and Paterson, D. G. Simplification of Flesch reading ease formula. Journal of Applied Psychology, October, 1951, 35, 333-37.

Flesch, R. F. Marks of Readable Style: A Study in Adult Education. New York: Bureau of Publications, Teachers College, Columbia University, 1943.

Flesch, R. F. Measuring the level of abstraction, Journal of Applied Psychology. December, 1950, 34, 384-90.

Forbes, F. W. and Cottle, W. C. A new method for determining readability of standardized tests. Journal of Applied Psychology, June, 1953, 37, 185-190.

Fry, Edward B. A readability formula that saves time. Journal of Reading, April, 1968, 513-516 and 575-578.

Gallant, R. Use of cloze tests as a measure of readability in the primary grades. Proceedings of the International Reading Association Convention, 1965, 10, 286-287.

Gray, W. S. and Leary, B. E. What Makes a Book Readable ...: An Initial Study. Chicago: The University of Chicago Press, 1935.

Guilford, J. P. Psychometric Methods, (second edition). New York: McGraw-Hill, 1954.

Gunning, R. The Technique of Clear Writing. New York: McGraw-Hill, 1952.

Horn, M. D. A Study of the Vocabulary of Children Before Entering the First Grade. Washington, D.C.: International Kindergarten Union, 1928.

Jenkinson, M. E. Selected processes and difficulties in reading comprehension. Unpublished doctoral dissertation, University of Chicago, 1957.

Johnson, G. R. An objective method of determining reading difficulty. Journal of Educational Research, April, 1930, 21, 283-87.

Klare, George R. The Measurement of Readability. Ames, Iowa: Iowa State University Press, 1963.

Klare, George R. Assessing readability. Reading Research Quarterly, 1974-75, 10 (1), 62-102.

Lewerenz, A. S. A vocabulary grade placement formula. Journal of Experimental Education, 1935, 3, 236.

Lewerenz, A. S. Vocabulary grade placement of typical newspaper content. Educational Research Bulletin, Los Angeles City Schools, September, 1930, 10, 4-6.

Lively, B. A. and Pressey, S. L. A method for measuring the "vocabulary burden" of textbooks. Educational Administration and Supervision, October, 1923, 9, 389-398.

Logan, Lillian M. and Logan, Virgil G. A Dynamic Approach to Language Arts. Toronto: McGraw-Hill of Canada, 1967.

Lorge, I. Predicting reading difficulty of selections for children. Elementary English Review, October, 1939, 16, 229-33.

Lorge, I. Predicting readability. Teachers College Record, March, 1944, 45, 404-19.

Lorge, I. The Lorge and Flesch readability formulae: a correction. School and Society, February 21, 1948, 67, 141-142.

McCall, W. A. and Crabbs, L. M. McCall-Crabbs Standard Test Lessons in Reading. New York: Bureau of Publications, Teachers College, Columbia University, 1925.

McCall, W. A. and Crabbs, L. M. Standard Test Lessons in Reading: Teacher's Manual for all Books. New York: Bureau of Publications, Teachers College, Columbia University, 1926.

McClusky, H. Y. A quantitative analysis of the difficulty of reading materials. Journal of Educational Research, December, 1934, 28, 276-82.

Neumeyer, Peter F. A structural approach to the study of literature for children. Reprint Number 9. Cambridge: Harvard R & D Center on Educational Differences, 1968.

Nice, M. M. An analysis of the conversation of children and adults. Child Development, 1932, 3, 240-246.

Ojemann, R. H. The reading ability of parents and factors associated with reading difficulty of parent education materials. University of Iowa Studies in Child Welfare, 1934, 8, 11-32.

Potter, T. C. A Taxonomy of Cloze Research, Part I: Readability and Reading Comprehension. Inglewood, California: Southwest Regional Laboratory for Educational Research and Development, 1968.

Rankin, E. F. An evaluation of the cloze procedure as a technique for measuring reading comprehension. Unpublished doctoral dissertation, University of Michigan, 1957.

Rankin, E. F. The cloze procedure: its validity and utility. In National Reading Conference Starting and Improving College Reading Programs; 8th yearbook, April, 1959.

Rankin, E. F. The cloze procedure: a survey of research. 14th Yearbook, National Reading Conference, 1965, 133-148.

Rinsland, H. D. A Basic Vocabulary of Elementary School Children. New York: Macmillan, 1945.

Rodgers, T. S. Linguistic consideration in the design of the Stanford Computer-based curriculum in initial reading. Technical Report No. 111. Stanford: Institute for Mathematical Studies in the Social Sciences, Stanford University, 1967.

Ruddell, R. B. The effect of oral and written patterns of language structure on reading comprehension. Unpublished doctoral dissertation, University of Indiana, 1963.

Spache, G. A new readability formula for primary-grade reading materials. Elementary School Journal, March, 1953, 53, 410-13.

Spache, G. Good Reading for Poor Readers, (Revised 9th Edition). Champaign, Illinois: Garrard, 1974.

Stone, C. R. Measures of simplicity and beginning texts in reading. Journal of Educational Research, February, 1938, 31, 447-50

Taylor, W. L. "Cloze Procedure": A new tool for measuring readability. Journal Quarterly, 1953, 30, 415-433.

Taylor, W. L. Recent developments in the use of "cloze procedure." Journal Quarterly, 1956, 33, 42-48, 99.

Thorndike, E. L. Reading as reasoning: a study of mistakes in paragraph reading. Journal of Educational Psychology, 1917, 8, 323-332.

Thorndike, E. L. The Teacher's Word Book. New York: Teachers College, Columbia University, 1921.

Tribe, E. B. A readability formula for the elementary school based upon the Rinsland vocabulary. Unpublished doctoral dissertation, University of Oklahoma, 1956.

Vogel, M. and Washburne, C. An objective method of determining grade placement of children's reading material. Elementary School Journal, 1928, 28, 373-81.

Washburne, C. and Morphett, M. V. Grade placement of children's books. Elementary School Journal, January, 1938, 38, 355-64.

Wert, J. E., Neidt, Charles O., and Ahmann, S. J. Statistical Methods in Educational and Psychological Research. New York: Appleton-Century-Crofts, 1954.

Wheeler, L. R. and Smith, E. H. A practical readability formula for the classroom teacher in the primary grades. Elementary English, November, 1954, 31, 397-99.

Appendix A

Judges' Summed Ranks on All 80 Books
and Scores Assigned by Regression Formulae

Formulative or Validative Set	Accession Number	Title	Author	Judges' Summed Ranks[1]	Regression Formula S[2]	Regression Formula L[3]
F	3-97	Ten Apples Up On Top	Theo Le Sieg	7	200.4	142.7
V	2-45	Hop on Pop	Dr. Seuss	8	267.1	267.5
F	4-99	Go, Dog, Go!	P.D. Eastman	15	206.2	160.5
F	2-44	Are You My Mother?	P.D. Eastman	25	379.4	291.0
V	3-159	Put Me In The Zoo	R. Lopshire	27	364.1	273.6
V	1-83	Come and Have Fun	E. Hurd	30	265.8	190.5
F	1-31	Green Eggs and Ham	Dr. Seuss	32	359.3	297.6
V	1-33	Who Will Be My Friends?	Syd Hoff	38	468.6	425.8
V	2-26	Where is Everybody?	Remy Charlip	44	158.7	168.3
F	4-113	Nobody Listens to Andrew	Guilfoile	53	311.2	315.3
F	4-41	Hector Protector	Maurice Sendak	56	186.1	367.6
F	3-64	Shoes for Angela	Ellen Snavely	67	591.3	483.2
F	3-149	King, the Mice & the Cheese	Gurney	71	461.1	480.6
F	4-42	Grizzwold	Syd Hoff	72	378.6	449.1
V	1-37	Summer	Alice Low	76	364.3	297.4

[1] Range = 7-391

[2] Range = 158-1881, Error of estimate 67.68

[3] Range = 142-1663, Error of estimate 81.93

Appendix A (continued)

Formulative or Validative Set	Accession Number	Title	Author	Judges' Summed Ranks[1]	Regression Formula S[2]	Regression Formula L[3]
V-2-30		I Should Have Stayed in Bed!	Lexau	79	321.6	270.1
V-2-43		The Bike Lesson	Berenstain	90	260.0	191.4
F-3-158		Oliver	Syd Hoff	90	297.5	351.6
V-1-19		Here Comes The Strikeout	Kessler	93	435.3	343.4
F-3-94		Little Bear	Else Minarik	99	358.3	279.9
F-4-84		Barefoot Boy	Gloria Miklowitz	100	651.3	555.4
V-3-22		Snowy Day	Ezra Keats	109	1007.	1114.
F-2-16		Red Fox and His Canoe	H. Benchley	117	446.5	321.7
F-2-23		Whistle for Willie	Ezra Keats	123	802.3	859.6
F-1-96		What Spot?	Crosby Bonsall	135	469.1	330.1
F-1-40		Case of the Hungry Stranger	Crosby Bonsall	135	438.3	359.4
V-2-4		The Case of the Cat's Meow	Crosby Bonsall	136	466.6	344.7
F-4-63		If It Weren't for You	Charlotte Zolotow	137	837.0	1044.
F-4-38		May I Bring a Friend	de Regniers	142	341.5	303.6
F-4-76		One Fish, Two Fish	Dr. Seuss	144	373.7	239.6

[1] Range = 7 - 391

[2] Range = 158-1881, Error of estimate 67.68

[3] Range = 142-1663, Error of estimate 81.98

Appendix A (continued)

Formulative or Validative Set	Accession Number	Title	Author	Judges'[1] Summed Rank	Regression[2] Formula S	Regression[3] Formula L
V-3- 2		The Cat In The Hat	Dr. Seuss	145	541.0	312.8
V-1-18		Little Bear's Visit	Else Minarik	151	473.2	416.8
F-4-69		Mud, Mud, Mud	Leonore Klein	168	616.2	682.2
V-2-46		Just Me	Marie H. Ets	169	468.9	402.1
F-3-12		Popcorn Dragon	Jane Thayer	174	717.3	625.0
V-3-29		Fox in Socks	Dr. Seuss	184	289.3	213.5
F-1-14		Let's Get Turtles	Millicent Selsam	186	567.2	350.0
V-3-11		Bedtime For Francis	Russell Hoban	189	729.5	466.3
F-2- 6		Greg's Microscope	Millicent Selsam	192	552.4	379.5
F-1-49		Blueberries for Sal	R. McCloskey	202	897.1	845.6
F-4-71		Mr. Bear Goes to Boston	Marion French	207	654.3	485.9
V-1-79		Shhh...Bang	Margaret Brown	211	446.1	558.3
F-4-88		The Three Robbers	Tomi Ungerer	215	604.8	807.5
F-1-55		Mike Mulligan	Burton	229	737.0	865.2
V-3-151		Keep Your Mouth Closed Dear	Aliki	231	778.7	810.9
F-2-59		White Snow, Bright Snow	Alvin Tresselt	233	716.0	903.6

[1]Range = 7 - 391

[2]Range = 158-1881, Error of estimate 67.68

[3]Range = 142-1663, Error of estimate 81.98

Appendix A (continued)

Formulative or Validative Set Accession Number	Title	Author	Judges'[1] Summed Ranks	Regression[2] Formula S	Regression[3] Formula L
F-[illegible]-56	Make Way For Ducklings	McClosky	235	799.9	763.9
F-3-112	Madeline's Rescue	L. Bemelmans	235	478.7	700.1
V-3-95	Zoo, Where Are You	Ann McGovern	240	789.7	601.5
V-1-13	Mississippi Possum	Miska Miles	247	879.3	755.8
F-4-154	Horton Hatches The Egg	Dr. Seuss	261	757.4	658.1
F-4-142	Frederick	Leo Lionni	263	597.6	687.8
V-3-105	Where The Wild Things Are	Maurice Sendak	265	1444.	1545.
V-2-125	One Morning In Maine	McCloskey	278	985.5	821.7
F-4-101	Lazy Tommy Pumpkin Head	Wm. DuBois	278	862.0	930.6
F-4-133	Martha the Movie Mouse	Arnold Lobel	287	704.4	650.2
F-4-157	The Moon Man	Tomi Ungerer	289	562.6	967.9
V-2-67	Rolling Round	Rolf Miller	292	933.2	1190.
F-1-47	Anatole and the Robot	Eve Titus	297	910.8	1566.
V-3-127	Baron Brandy's Boots	Peter Hughes	297	1113.	927.0
F-1-50	Yertle the Turtle	Dr. Seuss	298	723.5	434.9

[1]Range = 7-391

[2]Range = 158-1881, Error of estimate 67.68

[3]Range = 142-1663, Error of estimate 81.98

APPENDIX B

Protocol for Administering Placement Test and Reading Selections

1. E introduces student to task by saying, "We are building a library for children. We have some books and we hope you will tell us how you like them. We'd like you to read part of each one of them to us and then tell us what you think of them. Other children in your class will be helping us, too. At the same time we might find out if children your age can read these books easily or if some of them are too hard."

2. Record child's name, age, grade, class, the date, and the identification number from the tape cassette onto the form provided.

3. Explain to S how the notebook was made and show him a page from one of the real books and the copy of that same page in the notebook. Explain that there are words in the stories that are covered over or left out. Show him the first item in the demonstration story but do not allow him to read the page. Tell S that when he reads the story aloud he should say out loud the word that he thinks would go in the blank, that it is a guessing game, and that he will probably be able to tell what the covered word is by what has gone on in the story. Have S read from the selection until he comes to the first blank. (The text here reads, "Daddy said, 'wait _____. I must...' If the child hesitates, say, "One word is missing - it's not there. What word do you think should go there?" If S doesn't answer, ask, "Who would Daddy say 'wait' to?" If S says the correct word, E should reinforce with "wait Andrew - good." If S misses the word, E should point to the text and say, "Try Andrew" and have S reread. Follow with "good." This procedure may be repeated until cloze item #4. E then should say, "I won't help you anymore now. So you try

the rest by yourself. At the end of page 10, E reinforces S's performance with "Yes, that's good."

The first time that S is unable to read one of the words in the text and spends an inordinately long time over it, tell him that there may be words in this story, and in other stories too, that he does not know and cannot figure out. Make it clear that he should try but that he should not be upset if he cannot read a word. Assure him that he can just skip it and go on with the story. Explain that you will be unable to talk with him at all once he gets started on a selection.

Discuss the story with him after he finishes - for fun, don't make a test out of the discussion.

4. S must correctly guess three out of the last five cloze items in the demonstration story if he is to begin other selections in the test. If S misses more than two out of the last five cloze items, return him to class. Note: do not prompt words in the sample selection, other than the first four cloze items.

5. Begin your first S on Form A of the Placement Test, the second S on Form B, third S on Form A, etc., so that both forms will be used for the same number of children. To begin after the demonstration, say, "Now you go ahead and do it by yourself, I won't help you anymore. When a word is missing, you say the missing word. Remember only one word goes there and I can't tell you what the word is. If you don't know it, take a guess and go on reading. Remember only one word is missing each time. I can't tell you any of the missing words or what any of the words in the story are. Go ahead now. Start here." Indicate title. E moves his chair away from and in back of S and turns tape recorder on.

Places where S is to be reinforced verbally by E are marked with a check in E's copy of the text. They follow closely after either every third or fourth cloze item.

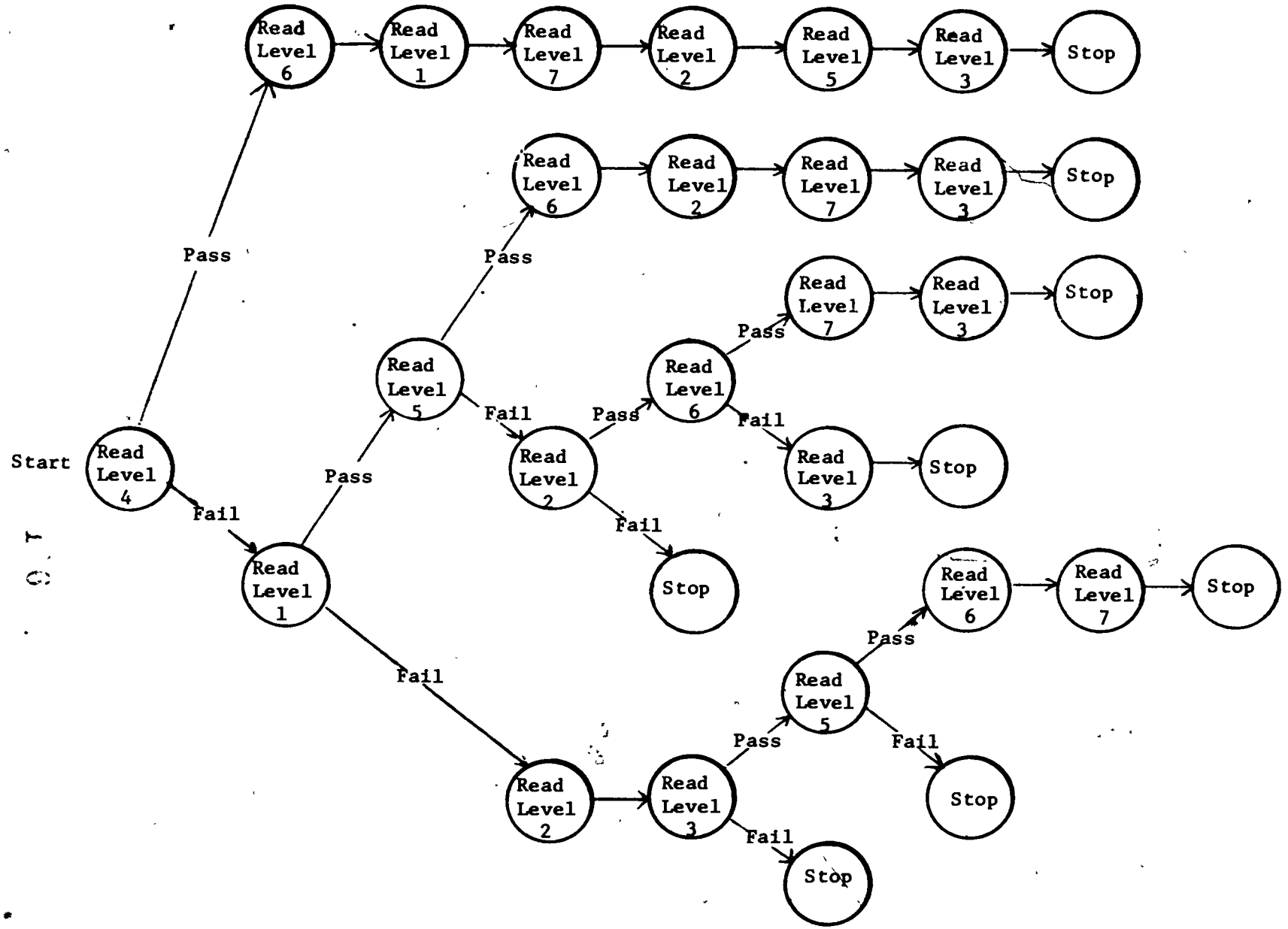
6. The only criterion of "Pass" or "Fail" in the readings from Form A or B is a five minute time limit. If S has not finished at least half of the selection (denoted by a line under the half-way word) in five minutes, E says, "That's all we'll read of this story today." E turns to the next selection for S following the "fail" arrow in Figure 2.3. If the selections all rate a "pass," they will be found in order in the notebook.

7. No S is to be kept out of class for more than 20 minutes at a time. E is not to correct either incorrect cloze items or mispronounced, misread words.

8. After S has finished a selection, E asks for comments on the book, e.g., would it be a good book to have in a library for children? Do you think other children would like to read it, too? Try to have some interaction about the content of the passage and information on S's likes and dislikes after three or four selections.

Also ask if S has read any of these books before or if anyone else has read them to him. Note on his record sheet if child has read book previously.

9. Each E should follow through with the book readings on the children he starts on the placement test insofar as this is possible.



Placement Test: Reading Selection Level Administration Sequence

Reading Selections for Placement Tests

The demonstration item was based upon Nobody Listens to Andrew by E. Guilfoile, and the passages making up test forms A and B are taken from books listed below.

Form A

Come and Have Fun by Edith Hurd (book 1, level 1)
Green Eggs and Ham by Dr. Seuss (book 2, level 3)
What Spot? by Crosby Bonsall (book 3, level 6)
Let's Get Turtles by Millicent Selsam (book 4, level 10)
Mississippi Possum by Miska Miles (book 5, level 13)
Orlando by Toni Ungerer (book 6, level 16)
Charlie and the Chocolate Factory by Raold Dahl (book 7, level 20)

Form B

Hop on Pop by Dr. Seuss (book 1, level 1)
Are You My Mother? by P. D. Eastman (book 2, level 3)
Red Fox and His Canoe by Nathaniel Benchley (book 3, level 6)
Greg's Microscope by Millicent Selsam (book 4, level 10)
One Morning in Maine by Robert McCloskey (book 5, level 13)
Rolling Round by Rolf Myller (book 6, level 17)
Paddington by Michael Bond (book 7, level 20)

APPENDIX C

Procedure for Finding Range of Reading Levels for Each Student to Read

1. Compute the $\bar{x}$ error score on Placement Test (included are levels 1, 3, 6, 10, 16, or 17, and 20).

2. Compute the SD on these same scores.

3. Add 1 SD to the $\bar{x}$ yielding a max. error score.

4. Assign as the "top book" for S that level where he obtained a score at or above his max. error score. Count down nine and assign this level as his "easiest book," thus spanning 10 levels.

   Example 1

Level	Student's Score
1	5.0
3	8.0
6	10.0
10	22.0
13	15.0
16	23.0
20	23.0

$\bar{x}$ = 15.14
SD = 7.64

max. error score = 22.78

Assign level 16 and "top book" and, counting down nine, level 7 as "easiest book." Range = levels 7 - 16.

5. Adjustments

   a. If S obtains a max. error score at a certain level in the Placement Test and then does not do so on a more difficult level of the test, do not use this lower level as S's "top book." Rather, proceed to the next instance of a max. error score and designate that level "top book." That is, S's "top book" must be where the max. error score is not immediately followed by a lower (better) score on a harder selection.

      For instance, if maximum error score = 12.5 and levels 6, 10, and 13 scores = 12.5, 8.0, and 14.0 respectively, you would assign level 13 as "top book" (not level 6), because level 10 was easier than level 6 for this S.

b. If the score that should determine selection of S's "top book" (as found in 4 above) is substantially higher than his max. error score (e.g., approximately another .5 SD) do not assign this level. Rather, move down to a point between this level and the next lower level on the Placement Test (see Example 2).

Example 2

Level	Student's Score
1	5.6
3	13.0
6	11.5
10	12.6
13	15.0
16	12.6
20	20.0

$\bar{x} = 12.9$
SD = 4.29

max. error score = 17.19

max. error score + .5(SD) = 17.19 + .5(4.29) = 17.19 + 2.15 = 19.34. Therefor do not assign level 20 (error score = 20.0) as "top book" because the score of 20.0 is substantially higher than his max. error score. Do move down several levels on the continuum of 20 books to, say, level 18 as "top book."

c. If "top book" is determined to be a level below level 10, assign only books from that point downward; do not move up in order to include 10 levels.

d. If "top book" is determined to be level 19 in the set, assign level 20 in addition to other ten books.

APPENDIX D

Cloze Deletion Rules

1. Mark every eighth word starting from the first word of the sample.
2. All words except preposition and conjunctions may be used as items.
3. When the eighth word is either a preposition or a conjunction, the word immediately to the left should be considered. If that word cannot be used, the word just to the right of the eighth word is to be considered. If that one is not acceptable, the word that is two words to the left is to be considered, then the one two words to the right, and so on until four words on each side of the eighth word have been considered. If no acceptable item according to the rules is found within these nine words, no item shall be chosen at that point in the text.
4. A few books have sentence structures such that the deletion of every eighth word would result in deleting the same word many times. In these cases a coin shall be tossed with heads equal to seven words and tails to nine. The rules above for determining the item should be used, substituting the word thus chosen (the seventh or ninth) for the eighth.

The scoring procedure was adopted of counting as correct only those items for which the exact word is replaced. A full discussion of why this scoring method was chosen over others is included in the section on scoring of the placement test (p. 31).

APPENDIX E

Important Aspects of Directions to Judges for Rank-Ordering

Take from the shelf the group of ten books listed on your ranking sheet, shelved according to their identification numbers. (These numbers have nothing whatsoever to do with the order of difficulty of the books.) First skim through the books, putting them in some rough order of difficulty. Then begin to order the books by reading first one book and then reading through another to compare the two. Lay them out on the table in order of difficulty and read another book. Read only as much of each book as you feel is necessary, but be sure to sample the text throughout the book. Place this third book in relation to the other two and continue in this manner, placing each book in the sequence between one that is easier and one that is more difficult than it. It is difficult to remember all ten books accurately and it will probably become necessary to re-examine some of the books as you go along. Re-examine those that you expect will surround the book you are working with and continue to place each into the sequence until all ten are ranked. Finally, go back over the set, skimming the books once again, making adjustments until you are satisfied with the resulting order.

The discriminations you are being asked to make are difficult ones, but no ties are allowed. If it is really impossible to decide, then make an arbitrary assignment, but this is to be considered a last resort and not at all advisable.

Other problems will arise when you are ranking books that are peculiar in some way: those written in poetry, for instance, or those dealing with extreme fantasy, those obviously translated from a foreign language or containing many foreign names and words, etc. Do the best

that you can with these, but do not communicate with the other judges about the problems that do arise. An underlying assumption for our procedure is that all of you have had the common base of the introductory training session and that further inter-communication will destroy that commonality.

APPENDIX F

Expansion of Dale List of 769 Easy Words*

Our rules for expanding the Dale list of 769 easy words followed those given by Dale, E. and Chall, J. in Educational Research Bulletin, February, 1948, 27, pp. 37-54, with the following exceptions made to give what appeared to be better distinctions between certain easy and different words.

a. Words that occur as both nouns and verbs in the language but which are infrequent as verbs were treated as nouns only, e.g., dog gave dogs and dog's as "easy" but not dogged or dogging.

b. Plural possessive nouns and inanimate possessive nouns were not included except:

1. boat's, train's, ship's (because, although inanimate, they are often personalized in children's books and so are included as "familiar")
2. Units of time: today's, yesterday's, year's, tomorrow's, days's, evening's, month's, night's, night's are also included
3. Possessives of groups of individuals are included: crowd's, company's, town's, family's, people's

c. Comparative and superlative forms of the following adjectives are included as "familiar" although the correctness of using these words in text might be questioned:

true, round, straight, blind, square; giving truer, truest, etc.

d. The following adverbs which could be formed grammatically are not included:

blackly, whitely, bluely, motherly, kingly, gamely, neighborly, sisterly

e. The following words which change meaning when put into adverbial forms are included as "familiar":

hardly, justly, likely, lively

f. ours, theirs, yours, and hers are included as "familiar"

*Rules are given in Dale and Chall, 1948.

Appendix G

Correlations of the Frequency of Common Words with Book-Difficulty as Determined by Ss Cloze Scores on Books in Sets I and II

	Set I	Set II
a) High Positive Correlation		
you	.40	-.29
very	.45	-.01
with	.53	.58
why	.40	.13
well	.54	.26
was	.49	.41
took	.40	.31
through	.63	.18
that's	.51	.14
over	.70	.24
or	.44	.10
one	.37	.06
off	.40	.16
of	.57	.60
now	.50	-.30
never	.49	.43
in	.47	.38
if	.57	.09
him	.42	.01
had	.59	.44
going	.38	-.26
get	.35	-.56
gave	.62	.21
from	.64	.66
for	.62	.30
could	.42	-.28
can't	.45	.12
called	.39	-.01
but	.54	.10
before	.53	.02
been	.45	.45
another	.38	.28
and	.49	.58
after	.37	-.38
about	.50	.41
a	.53	-.13
are	.30	-.60
because	.31	.29
how	.31	-.13
I'll	.32	.04
more	.34	.17
to	.36	-.07
the	.31	.19
then	.30	-.13
when	.34	.21

	Set I	Set II
b) High Negative Correlation		
your	-.32	-.27
will	-.46	-.41
who	-.45	.01
went	-.48	-.72
we	-.36	-.34
sat	-.42	-.43
said	-.47	-.25
my	-.39	-.31
good	-.46	-.13
fast	-.49	-.44
come	-.43	-.03
can	-.40	-.29
away	-.3	.09
asked	-.3	.14
am	-.45	-.51
be	-.30	-.26
did	-.34	-.54
have	-.33	-.25
out	-.38	-.35
ran	-.33	-.29
yes	-.31	-.42
us	-.25	.24
things	-.25	.10

p<	.10	.05	.025	.01	(two tailed test)
r=	.378	.444	.516	.561	

APPENDIX H

Symbols and Rules Used for Key-Punching Text of Books

FOR	USE
Quotation marks	Two apostrophes
Question mark	Dollar sign
Exclamation point	Asterisk
Comma	Comma
Semi-colon	Equal sign
Colon	Plus sign
Parentheses, ellipses, and dashes	Three hyphens

RULES:

Leave a space before and after all above symbols

Omit periods after abbreviations: Mr and Mrs not Mr. and Mrs.

At the end of a page: 1) Use / where there is terminal punctuation
2) Use // where there is no terminal punctuation

Punch to end of word closest to column 80. Do not hyphenate words not hyphenated in book.

If quotation marks appear at the beginning of a sentence, but do not appear at the end, then punch two apostrophes at both the begining and at the point where the quotation ends and the quotation marks should appear.

Punch text exactly as it appears except for the above substitutions, additions, and deletions.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20----------
R P 0 0 1 X X X X 2 0 X C 5 1 X TEXT AS WRITTEN IN BOOK

Consecutively from 001 to number of cards used for each book.

Level number

Story accession number

X=blank space

Appendix I

"Concepts" Requested on Computer Analysis of the Text of Each Book

Personal-personal Pronouns

I	mine	we	you	your
I'd	my	we're	you're	yourself
I'll	myself	we'd	you'd	yourselves
I'm	you	we'll	you'll	
I've	yourselves	we've	you've	
me	us	ours	yours	

Third-person Personal Pronouns

he	himself	hers	they've
he's	she	herself	them
he'd	she's	they	their
he'll	she'd	they're	theirs
him	she'll	they'd	themselves
his	her	they'll	

Personal Pronouns

(Total of count of Personal-personal Pronouns and Third-person Personal Pronouns above)

Impersonal Pronouns

all	itself	those	whoever
any	many	who	whomever
both	most	who's	whosever
either	much	who'd	whichever
each	neither	who've	
few	none	who'll	
it	several	who're	
it'd	some	which	
it'll	that	whom	
it's	these	whose	
its	this	that's	

Pronouns

(Total count of three categories of pronouns above)

Appendix I (continued)

Prepositions

aboard	below	in	throughout
about	beneath	inside	to
above	beside	into	toward
across	between	like	under
after	beyond	near	until
against	by	of	unto
along	concerning	off	up
among	despite	on	upon
around	down	onto	with
at	during	over	within
before	for	since	without
behind	from	through	

Conjunctions

and	for	or	yet
but	nor	so	

Dale Words

2000 words derived from Dale list of 769 words (available upon request from Popp and Porter)

Words

(Count of every word in text)

Sentences

(Count of periods, question marks, exclamation points)

Pages

(Count of each page in book containing printed text)

Sentences carried over from one page to another

(Coded by double slashes)

Intercorrelations Among Criterion (Judges' Summed Ranks) and Textual Variables (Formative Set, N = 50)

Variable	2	3	4	5	6	7	8	9	10
1. Words 7+ letters/word	.889**	-.660**	.176	.251	.261	-.006	.201	-.414**	-.175
2. Words 7+ letters/sentence		-.588**	.504**	.589**	.595**	.111	.543**	-.397**	-.015
3. Dale 769 words/word			-.192	-.301*	-.271	-.016	-.244	.493**	.158
4. Dale 769 words/sentence				.984**	.958**	.324*	.834**	-.065	.613**
5. Words/sentence					.971**	.317*	.866**	-.128	.550**
6. Prepositions/sentence						.330*	.808**	-.200	.460**
7. Conjunction/sentence							.235	-.177	.058
8. , : ;/sentence								-.029	.472**
9. Pers. pers. prn./word									.649**
10. Pers. pers. prns./sentence									
11. Pers. prns./word									
12. Pers. prns./sentence									
13. Imp. + 3rd pers. prns./word									
14. Imp. + 3rd pers. pers. prns./sentence									
15. Imp. prns./pers. prns.									
16. Imp. prns./pers. pers. prns.									
17. Imp. + 3rd pers. prns./pers. pers. prns.									
18. Pers. prns./conjunction									
19. Pers. pers. prns./conjunction									
20. Words/page									
21. Number words in book									
22. Number pages in book									
23. Carried-over sentences/page									
24. Judges' summed ranks									

$p <$	.05	.01	(two-tailed test)
$r =$	.282	.365	

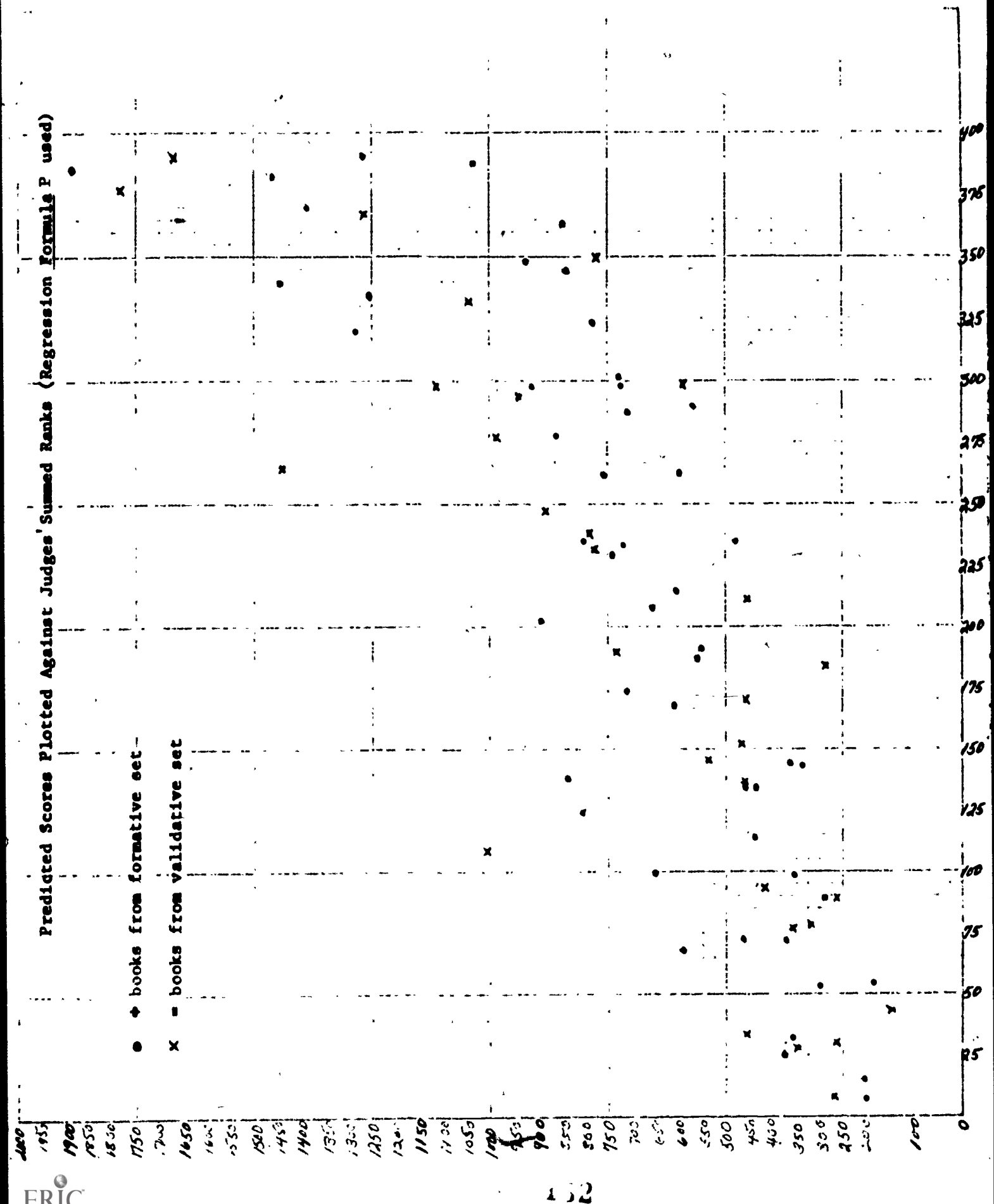
11	12	13	14	15	16	17	18	19	20	21	22	23	24
335	.007	.046	.320*	-.085	.196	.235	-.197	-.343*	.333*	.209	-.319*	.167	.516**
403**	.242	-.040	.582**	.012	.383**	.387**	-.327*	-.441**	.345*	.276	-.299*	.217	.645**
413**	.030	-.138	-.383**	.085	-.309*	-.296*	.191	.395**	-.237	-.137	.389**	-.167	-.545**
212	.827**	-.201	.750**	.015	.279	.267	-.414**	-.411**	.158	.199	-.238	.553**	.422**
260	.791**	-.167	.792**	.029	.349*	.325*	-.402**	-.429**	.180	.200	-.268	.489**	.499**
278	.745**	-.098	.828**	.012	.379**	.376**	-.403**	-.471**	.211	.208	-.309*	.461**	.523**
101	.236	.135	.393**	-.121	.129	.184	-.092	-.234	.220	.238	-.122	.222	.272
250	.626**	-.255	.630**	.122	.423**	.347*	-.351*	-.285*	.158	.168	-.197	.379**	.431**
767**	.354*	-.479**	-.480**	-.309*	-.525**	-.530**	.305*	.683**	-.068	.020	.287*	.172	-.392**
371**	.869**	-.433**	.100	-.141	-.306*	-.312*	-.038	.204	.022	.087	.026	.578**	-.077
	.306*	.084	-.282*	-.626**	-.485**	-.[illegible]**	.446**	.590**	-.067	-.027	.179	.047	-.393**
		-.133	.536**	-.261	-.093	-.073	-.150	-.074	.140	.184	-.144	.606**	.192
			.412**	.004	.098	.117	.067	-.247	.116	.051	-.098	-.186	.147
				.052	.437**	.417**	-.392**	-.597**	.299*	.285*	-.320*	.299*	.632**
					.248	.102	-.379**	-.270	.036	.045	.112	-.086	.100
						.958**	-.307*	-.424**	-.099	-.177	-.312*	-.204	.219
							-.277	-.424**	-.131	-.210	-.353*	-.195	.195
								.845**	-.193	-.114	.424**	-.271	-.393**
									-.190	-.083	.481**	-.169	-.526**
										.831**	-.142	.456**	.685**
											.259	.460**	.585**
												-.079	-.295*
													.260

Intercorrelations Among Criterion (Judges' Summed Ranks) and Textual Variables (Validative Set, N = 30)

Variable	2	3	4	5	6	7	8	9	10
1. Words 7+ letters/word	.937**	-.372*	.402	.476**	.457*	.346	.690**	-.467**	-.220
2. Words 7+ letters/sentence			.653**	.722**	.695**	.412*	.667**	-.496**	-.205
3. Dale 769 words/word			.055	-.106	-.065	-.105	-.524**	.481**	.380*
4. Dale 769 words/sentence				.986**	.969**	.324	.206	-.411	-.117
5. Words/sentence					.978**	.348	.311	-.474**	-.165
6. Prepositions/sentence						.284	.244	-.465**	-.165
7. Conjunction/sentence							.361	.051	.271
8. , : ;/sentence								-.303	-.109
9. Pers. pers. prn./word									.900**
10. Pers. pers. prns./sentence									
11. Pers. prns./word									
12. Pers. prns./sentence									
13. Imp. + 3rd pers. prns./word									
14. Imp. + 3rd pers. pers. prns./sentence									
15. Imp. prns./pers. prns.									
16. Imp. prns./pers. pers. prns.									
17. Imp. + 3rd pers. prns./pers. pers. prns.									
18. Pers. prns./conjunction									
19. Pers. pers. prns./conjunction									
20. Words/page									
21. Number words in book									
22. Number pages in book									
23. Carried-over sentences/page									
24. Judges' summed ranks									

$p < \frac{.05}{r = .361} \mid \frac{.01}{.463}$ (two-tailed test)

11	12	13	14	15	16	17	18	19	20	21	22	23	24
244	.274	.405*	.532**	-.142	.298	-.542**	-.006	-.259	.594**	.442*	-.558**	.334	.787**
277	.446*	.391*	.720**	-.105	.292	.559**	-.061	-.297	.566**	.429*	-.591**	.412*	.819**
419*	.139	-.174	-.173	.065	.048	-.122	.089	.282	-.383*	-.387*	-.028	-.160	-.503**
254	.651**	.183	.825**	.104	.161	.211	-.275	-.385*	.189	.144	-.510**	.408*	.513**
304	.638**	.219	.854**	.070	.145	.235	-.272	-.412*	.267	.222	-.507**	.434*	.603**
319	.599**	.183	.815**	.106	.174	.193	-.272	-.405	.290	.233	-.516**	.418*	.579**
055	.379*	-.024	.248	-.027	-.017	.304	-.189	-.154	.111	.108	-.149	.337	.319
063	.273	.409*	.425*	-.308	.090	.396*	.017	-.196	.625**	.621**	-.359	.365*	.765**
806**	.178	-.353	-.513**	-.336	-.307	-.363*	.385*	.700**	-.254	-.246	.266	-.142	-.437*
742**	.435*	-.345	-.300	-.373*	-.323	-.345	.284	.544**	-.084	-.110	.023	.058	-.123
	.511**	.212	-.098	-.676**	-.003	-.265	.338	.508**	-.205	-.202	.070	-.068	-.321
		.402*	.704**	-.554**	.162	-.032	-.020	-.028	.086	.049	-.387*	.367*	.327
			.674**	-.470**	.518**	.314	.052	-.206	.195	.195	-.187	.168	.253
				.235	.401*	.638**	-.175	-.393*	.269	.241	-.428*	.419*	.545**
					-.013	.129	-.156	-.155	.044	-.076	-.031	-.003	-.112
						.445**	.028	-.137	-.019	-.086	-.229	.127	-.032
							.258	.032	.268	.199	-.184	.281	.306
								.901**	.328	.214	.085	.186	-.095
									.090	.016	.243	.037	-.307
										.901**	-.365*	.641**	.745**
											-.067	.482**	.698**
												-.407*	-.500**
													.436*


Predicted Scores Plotted Against Judges' Summed Ranks (Regression Formula P used)
● = books from formative set
× = books from validative set
Predicted Scores

Appendix K (continued)